Structure, function, and mechanism of β-barrel protein assembly in *Campylobacter jejuni*

by

Matt Hardcastle

BEng MSc (Hons)

Thesis submitted to the University of Nottingham in part fulfilment of the degree of Doctor of Philosophy, March 2018

Abstract

Campylobacter jejuni is one of the world's leading causes of food-mediated bacterial infection, affecting millions of people every year. BamA, a highly conserved, essential membrane protein found in *C. jejuni* and other Gram-negative bacteria is responsible for folding a number of essential β-barrel outer membrane proteins making it vital for survival in this pathogenic organism. Along with essential roles in basic physiological functions, such as nutrient import and export, outer membrane proteins are also found in a number of multidrug resistance and virulence mechanisms in pathogenic organisms, making them and BamA popular targets for therapeutic intervention. BamA is well characterised in model organisms, however the conversed mechanism which facilitates assembly in different species remains unclear. In this work, the biogenesis of β-barrel proteins is investigated in *C. jeuni* NCTC-11168. A recombinant expression system is developed and used to express recombinant *Cj*BamA (*Cj*0129) in an *E. coli* BL21 host, followed by subsequent purification and biophysical characterisation. A pelB leader sequence is used to target the recombinant protein for outer membrane localisation, with an extended 10x-polyhistidine tag and purified by immobilised metal affinity chromatography. Secondary structure analysis is performed using synchrotron radiation circular dichroism confirming the presence of a dominant β-sheet fraction. Furthermore, bioinformatics tools are used to investigate the structure of BamA homology models and explore the structural relationships between BamA and a previously unidentified accessory BamD protein found in the *Campylobacter* genus.

Table of Contents

Abstract....3

Table of Contents....4

List of figures....8

Abbreviations....10

Chapter 1- Introduction

1.1 Campylobacter jejuni NCTC-11168....13

1.1.1 Clinical significance

1.1.2 Industrial relevance and antibiotic resistance

1.1.3 Campylobacter genetics and taxonomy

1.1.4 Bacterial membranes and taxonomy

1.1.5 Outer membrane structure and LPS

1.2 Outer membrane biogenesis....19

1.2.1 Inner membrane secretion

1.2.2 Lol

1.2.3 Bam

1.2.4 Structure of the β-Barrel Assembly Machinery

1.2.5 Mechanism of β-Barrel Assembly

1.2.6 β-barrel proteins (Omp85 superfamily) in other domains of life

1.2.7 Target substrate recognition

1.2.8 BamD and Auxiliary Lipoproteins

1.2.9 The Essential Nature of BamA

1.2.10 Species specificity in target recognition and folding

1.2.11 Dynamic behaviour of BamA

1.2.12 Outer membrane protein biogenesis in *C. jejuni*

1.3 Methodology....33

1.3.1 Aims and objectives

1.3.2 Research techniques

Chapter 2 - Materials and Methods

2.1 Genetic Engineering....37

2.1.1 Buffer preparation

2.1.2 Batch culture preparation and inoculation
2.1.3 Low temperature induction
2.1.4 Harvesting
2.1.5 Lysis
2.1.6 Outer Membrane Isolation
2.1.7 Crude Membrane Isolation
2.1.8 Immobilised Metal Affinity Chromatography
2.1.9 Size Exclusion Gel Permeation Chromatography
2.1.10 Protein Concentration Determination
2.1.11 Liposomal Reconstitution

Chapter 3. Membrane protein expression, purification, and characterisation
3.1 Introduction 41
3.1.1 Identification of BamA in *C. jejuni* NCTC-11168
3.2 Background and Theory 42
3.2.1 X-Ray Crystallography
3.2.2 Membrane protein extraction with detergent
3.2.3 Application of membrane mimetics
3.2.4 Thermodynamics of membrane protein solubilisation
3.2.5 Application of detergents in membrane protein preparation
3.2.6 Non-detergents membrane mimetics
3.2.7 Dynamic Light Scattering
3.2.8 Atomic Force Microscopy
3.3 Results 57
3.3.1 Particle size analysis by DLS and AFM
3.3.2 Cloning, low temperature expression and purification of BamA
3.3.3 Detergent buffer screening
3.3.4 Crystallisation and X-ray diffraction crystallography
3.4 Discussion 65
3.4.1 Affinity tag placement disrupts BamA folding
3.4.2 Localisation of *C. jejuni* BamA limits growth in *E. coli*
3.4.3 Yield optimisation for crystallisation
3.4.4 Application of high-end structural techniques

Chapter 4. Template based protein structure prediction of BamA and BamD proteins
4.1 Introduction 73
4.2 Background and theory 73
4.2.1 Evolution of Protein Structure and Complex Assembly Machinery

4.2.2 Theory of structural macromolecular evolution
4.2.3 Dynamic programming and template-based structure prediction
4.2.4 Protein homology modelling
4.2.5 Scoring and template fitting
4.3 Results 80
4.3.2 Co-evolutionary Signal Detection
4.3.3 Identification of BamD and structural investigation
4.4 Discussion 89
4.4.1 The Protein Folding Problem

Chapter 5. Secondary structure prediction, estimation and thermolytic degradation by synchrotron radiation circular dichroism
5.1 Introduction 96
5.2 Background and Theory 96
5.2.1 Secondary Structure Estimation of Protein
5.2.2 Protein Concentration Determination
5.2.3 CONTIN/LL
5.2.4 DSSP
5.3 Experimental Set-up 101
5.3.1 Thermolytic Degradation
5.4 Results 103
5.4.1 Secondary structure estimation
5.4.2 Thermolytic degradation and protein stability assay
5.5 Discussion 110

Chapter 6. Molecular dynamic simulations of BAM proteins
6.1 Introduction 112
6.2 Background and Theory 112
6.2.1 Theoretical considerations
6.2.2 Defining a System
6.2.3 Force Fields and Energy Functions
6.2.4 Bonded Terms
6.2.5 Non-Bonded Terms
6.3 Computational methods 117
6.3.1 System setup and the CHARMM-GUI
6.3.2 Trajectory file analysis
6.4 Results 120
6.4.1 Analysis of all-atom simulations of BamA
6.4.2 Trajectory analysis of *Cj*BamA and *Ng*BamA simulations

6.4.3 Extended P23 loop interactions with membrane
6.4.4 Extended simulation of *Cj*BamA
6.4.5 Analysis of internal domain dynamics
6.4.6 NMA, PCA and residue cross-correlation network analysis
6.4.7 Course-grained simulation of full-length BamA
6.4.8 Simulation of BamD models
6.5 Discussion 142

Chapter 7. Conclusions and Future Work

Acknowledgements 151

Bibliography 152

Supplementary material 164

List of Figures

Figure 1.1 Structure of the bacterial outer membrane....18
Figure 1.2 Schematic of outer membrane protein assembly by the BAM complex....20
Figure 1.3 Open and closed crystal structure poses off BamA....24
Figure 1.4 Subunits of BamA POTRA and β-barrel domains....27
Figure 1.5 Membrane distortions observed in molecular dynamics simulations....31
Figure 3.1 Schematic of X-Ray crystal diffraction....46
Figure 3.2 Chemical structures of detergent molecules....49
Figure 3.3 Membrane mimetics in membrane protein solubilisation and reconstitution....53
Figure 3.4 Schematic diagram of dynamic light scattering....54
Figure 3.5 Particle size analysis by dynamic light scattering of reconstituted BamA proteoliposomes....59
Figure 3.6 Atomic force microscopy image of BamA-DMPC liposomes....61
Figure 3.7 Western Blotting and SDS-PAGE of BamA expression....63
Figure 3.8 Western blotting and ultracentrifugation dispersity assay....66
Figure 3.9 False positive crystal and X-ray diffraction patterns....68
Figure 4.1 Models of *C. jejuni* BamA generated by MODELLER, PHYRE2 82
Figure 4.2 Structural alignment of *Cj*BamA model and template 4k3b....83
Figure 4.3 Contact map prediction of BamA by co-evolutionary signal detection....87
Figure 4.4 Secondary structure prediction of putative *C. jejuni* BamD....90
Figure 4.5 Models of *Cj*BamD generated using MODELLER, PHYRE2, and I-TASSER....91
Figure 5.1 Reference spectra for secondary structure estimation....99
Figure 5.2 Synchrotron radiation circular dichroism instrument rig....102
Figure 5.3 Secondary structure estimation of *Cj*BamA....105
Figure 5.4 BamA-DDM thermal melt folding profile....107
Figure 5.5 BamA-SB12 thermal melt folding profile....108
Figure 5.6 Heat modifiability of BamA....109
Figure 6.1 Flowchart of MD simulation stages in GROMACs....121
Figure 6.2 Root mean square fluctuation of *Cj*BamA....122
Figure 6.3 Root mean square fluctuation of *Ng*BamA....123
Figure 6.4 Three replicate simulations of *Cj*BamA and *Ng*BamA....126
Figure 6.5 BamA POTRA domain loop interacts with the bilayer....127
Figure 6.6 Extended run simulation of I-TASSER *Cj*BamA model in NAMD....129
Figure 6.7 Internal dynamics of BamA sub-domains....131
Figure 6.8 Normal mode analysis of *Cj*BamA....132
Figure 6.9 Principal component analysis of *Cj*BamA....133
Figure 6.10 Principal component analysis of *Ng*BamA....134
Figure 6.11 Residue cross correlation network of a *Cj*BamA simulation....135
Figure 6.12 Residue cross correlation network of a *Ng*BamA simulation....137
Figure 6.13 Root mean square fluctuation of all-atom and coarse-grained BamA....139
Figure 6.14 Collapse of coarse grained BamA model....140
Figure 6.15 GROMACs simulations of BamD models....141

Figure 7.1 Mechanism of OMP assembly by BamA and BamD.....147
Figure 7.2 Detrended trajectory of 10 ns BamA simulation.....149
Figure 7.3 Resulting harmonic frequencies of vibrations in a BamA simulation.....149

Abbreviations

AFM Atomic force microscopy
Bam β-barrel assembly system
BAM β-barrel assembly machinery complex
BamABCDE β-barrel Assembly Machinery units A to E
BLAST Basic Local Alignment Search Tool
CD Circular dichroism
CHAPS - 3-[3-(cholamidopropyl) dimethylammonio]-1-propanesulfonate
CHAPSO - 3-[3-(cholamidopropyl) dimethylammonio]-2-hydroxy-1-propanesulfonate
CHARMM Chemistry at HARvard Macromolecular Mechanics
CMC - Critical micelle concentration
DDM Dodecyl-β-D-maltoside
DMPC - 1,2-myristoyl-1-*sn*-glycero-3-phosphocholine
DOPE - 1,2-oleoyl-*sn*-glycero-3-phosphoethanolamine
EM – Electron microscopy
ExPASy – Swiss Institute of Bioinformatics Expert Protein Analysis System
GREMLIN – Generative Regularized ModeLs of proteINs
GROMACS –GROeninger Molecular MAChine
GUVs – Giant Unilammelar Vesicles
HMM – Hidden Markov models
LDAO – Lauryldimethylamine N-oxide
LOS – Lipooligosaccharide
LPS – Lipopolysaccharide
MD – Molecular dynamics
NMR – Nuclear magnetic resonance
OG - *n*-octyl-β-D-glucoside
OMP – Outer membrane protein
PAGE – Polyacrylamide gel electrophoresis
PCA – Principal component analysis
PCR – Polymerase chain reaction
PRANK - Probabilistic alignment kit
SB12 - Sulfobetaine-12
T_m – Melting temperature
TEV - Tobacco etch virus
TPR – Tetratricopeptide repeat

SI Units

kg = kilogram (mass)
m = meters (length)
s = seconds (time)
K = kelvin (temperature)
V = volts (electrical potential)

List of conversions

min = minutes (60 s)
hrs = hours (3600 s)
1°C = 273 Kelvin (K)
1 psi = 6895N/m^2

List of standard prefixes

n *nano*	=	1 x 10^{-9}
μ *micro*	=	1 x 10^{-6}
m *milli*	=	1 x 10^{-3}
c *centi*	=	1 x 10^{-2}
k *kilo*	=	1 x 10^{3}
M *mega*	=	1 x 10^{6}
G *giga*	=	1 x 10^{9}

Chapter 1- Introduction

All living organisms must undergo essential processes to survive. Most biochemical pathways are highly co-ordinated reactions, involving many different units, but some of the most complex processes occur in and around the area of biomembranes. These processes include import and export functions found at the interfaces between cells and their external environment which have evolved together over billions of years [1]. Many unique compounds take part in these reactions, broadly categorised into four main groups; carbohydrates, lipids, proteins, and nucleotides. The first two provide a source of energy and material to make other molecules, while lipids also act as structural components in membranes - stored in regular geometric assemblies to provide shape and stability to cells or the organelles within them.

1.1 Campylobacter jejuni NCTC-11168

C. jejuni is a Gram-negative, non-spore forming, micro-aerophilic bacterium, which means it requires oxygen to survive but at a lower concentration than normal in the Earth's atmosphere [2]. It is part of a diverse group of organisms called Campylobacters, which include species found in a number of different environmental niche. *C. jejuni* is particularly common in poultry and other avian species, normally harboured in the digestive tract where they usually survive without causing harm to the host animal, but they are also bacteria are also found in other mammals, reptiles and shellfish.

1.1.1 Clinical significance

In humans, *C. jejuni* is currently the leading cause of food-mediated bacterial gastroenteritis in the developed world, ahead of more well recognised organisms such as *Salmonella, E. coli* and *Listeria* [3]. Infection with *C. jejuni* typically causes hallmark symptoms of gastroenteritis, which include a mild fever, stomach cramps, nausea, and diarrhoea. Although it is rarely a direct cause of death, infections have led to fatalities in certain situations involving the use of prescription antibiotics [4]. *C. jejuni* responds to ciprofloxacin and other fluoroquinolone drugs, however the administration of antibiotics in certain cases of infection can be risky as further disruption of the gut flora can worsen diarrhoea, leading to a more serious state of dehydration [5]. Further post-infection complications also include the neuropathic Guillain-Barre syndrome - a serious condition which causes debilitating muscular paralysis for which is currently difficult to treat [6].

Most mechanisms of pathogenicity in *C. jejuni* and other Gram-negative pathogens are associated with the outer membrane. These include toxin secretion, multidrug resistance and efflux, host cell colonisation, as well as protection and evasion from the host immune system [7] [8]. *C. jejuni* can also develop a protective polysaccharide capsule, known to protect against human blood serum resistance and contribute towards invasion of intestinal epithelial cells in models of pathogenicity [9]. The lipooligosaccharide (LOS) molecules, which form the capsule, are highly irregular compounds - the structures of which have been shown to as resemble that of human gangliosides [10]. This mechanism of molecular mimicry is also known to play a role in the development of Guillain-Barre syndrome and another auto-immune diseases including Miller-Fischer syndrome, which have both been linked to *C. jejuni* infections [11].

1.1.2 Industrial relevance and antibiotic resistance

Although it is normally a commensal organism that does not cause harm in most animal hosts, bovine campylobacteriosis can occur in cattle, most commonly in young calves, leading to diarrhoea and significant weight loss [12]. Although this condition can be treated with antibiotics, it is better avoided with good animal husbandry and hygiene practices. Poor practises in the agricultural and food processing industries though have caused widespread contamination in products and on packaging, leading to exposure and potential infection in humans. Furthermore, antibiotic resistance has been shown to be almost synonymous with *C. jejuni* strains isolated from broiler chickens in factories both in Brazil and China [13] [14]. In one of these studies, almost 90% of those isolates tested encoded genes imparting resistance to ciprofloxacin and up to 80% displayed resistance to tetracyclin – two antibiotics that are commonly administered to treat infection in both humans and livestock.

It is widely accepted that the rise in resistance against antimicrobial and antibiotic compounds in Gram-negative bacteria poses a significant threat to human wellbeing [15]. For this reason, it is important to consider the ongoing use of antibiotics to treat infection, especially for animal livestock. Although *C. jejuni* infections do not normally warrant pharmacological intervention in humans, resistance mechanisms have still developed against macrolide and fluoroquinolone drugs including nalidixic acid, tetracyclin and erythromycin, are still prevalent in this species [16]. This phenomenon is thought to be largely avoidable though, influenced in part by the indiscriminate

use of antibiotics in chickens and other animals[17] which has now caused an unfortunate economic and ethical dilemma which urgently needs addressing.

1.1.3 Campylobacter genetics and taxonomy

The *C. jejuni* NCTC-11168 strain genome is approximately 1.6 Mb pairs long and codes for at least 1,572 different unique proteins, according to KEGG (Kyoto Encyclopaedia of Genes and Genomes). Meaning "curved bacteria", the Campylobacter genus is comprised of over 30 other unique species including many known human pathogens such as *C. coli*, *C. rectus* and *C. fetus*. The genus belongs to the ε-proteobacteria class which also includes other closely related organisms such as *Arcobacter*, *Helicobacter* and *Wolinella* which are also known pathogens which can infect humans. Notably, *H. pylori* has been estimated to be present in up to half the world's population, and is a common cause of chronic gastritis which can also lead to several other serious diseases of the gastrointestinal tract including gastric cancer and duodenal ulcers [18].

The genetic diversity found in this phylum can be staggering though. Evolution of bacteria is strongly influenced by horizontal gene transfer, a process by which genetic material is shared between sibling colonies. This makes it difficult to definitively determine taxonomic groups in bacteria and this task is especially problematic for this genus [19]. *C. jejuni* strains also contain hyper-variable sequences, that are matching genes with many unique instances found in the same species, making classification even more problematic [20]. Multi-locus sequence typing and whole genome sequencing have been applied to *C. jejuni* and *C. coli* to understand how population structure and phenotypic variations in these human pathogen species occurred. Recombination was also found to be a strong driving force in the evolution of Campylobacter species with large scale introgression occurring between *C. jejuni* and *C. coli*. Other interesting members of the ε-proteobacteria class have been isolated from deep-sea hydrothermal sulphur vents in North America and South-East Asia, eluding to origins in naturally hot and acidic environments. These species, such as *Sulfurovum lithotrophicum* and *Nautilia profundicola* likely share close evolutionary paths with *Campylobacter* species, yet over time each have diverged to suit their own environmental niche. Along with other mechanisms, such as frame shift mutations, which impart expressive genetic diversity, Campylobacters are regularly spread amongst various hosts making them a diverse set of organisms evolving through rapid adaptation to a complex set of environments.

1.1.4 Bacterial membranes and taxonomy

The ability to target and traffic molecules is critical for biological processes in all species. Within the context of the cell, membranes may appear to have a passive role compared to other components that diffuse around the cell, but they are in fact highly active, dynamic structures. Membranes provide structure and stability while protecting the inner components from a potentially harsh external environment, but they also form a part of diverse assemblies with various roles adapted to the cells or organelles they surround. As such, different membranes are composed of different types of lipids, suited for function in various roles and environments.

Membranes are also used as an important indicator when differentiating between species and the study of this, taxonomy, has been influenced early on by observations made based on the shape and size of membranes. Plants and animals can be easily distinguished from single-celled organisms by the cellular structure, or *morphology,* which arises from different compositions of lipid components. These two groups are found in the Eukaryotic domain, which is generally defined to encompass organisms whose cells have a nucleus and other membrane bound organelles [21]. In contrast to the Eukaryotes, Prokaryotes are comprised of two domains of unicellular organisms which lack a membrane bound nucleus, known as Bacteria and Archaea [22]. Their cellular structure is typically less complex inside the cell but the membranes which surround them are multifaceted, functional components and essential for survival. Bacterial membranes are therefore diverse and can adjust their composition in response to environmental changes such as temperature and pH, allowing them to protect the organism in unfavourable or extreme conditions [23].

Due to the vast diversity of Prokaryotes, bacteria are further divided into two distinct groups - Gram-negatives and Gram-positives - named after the Gram stain test that is used to determine the presence of exposed peptidoglycan. In Gram-positive organisms a crystal purple dye is absorbed and retained by a thick wall of peptidoglycan which surrounds the cell [24]. Instead of the thick, protective peptidoglycan layer, Gram-negatives have a double layered, inner and outer membrane system, the very outer leaflet of which is composed of a charged lipopolysaccharide layer that does not retain the dye. The generation of LPS and the peptidoglycan cell wall are both thought to be evolutionary traits which have been retained by organisms found in extreme or harsh conditions and contribute to the prevalence of bacteria in a wide range of environmental niches [25] [26].

Studying the bacterial membrane can also give further signs of a species origins and evolutionary path. Not all bacterial organisms can be easily categorised into the Gram-negative and Gram-positive groups though. One specific case has been discovered which reveals the limitations of the Gram staining test: the Negativicutes and Halanaerobiales. These organisms are part of a group which include a textbook example of bacteria with a *monoderm* (single membrane) -the Firmicutes – that includes notable pathogens such as *Staphylococcus*. Negativicutes and Halanaerobiales however were surprisingly found to have a *diderm* (double membrane), which includes LPS and outer membrane proteins and it is unclear how such a transition between the two occurred. Two current hypotheses have been explored in literature, described as *diderm first* [27] or *monoderm first* [28]. Although most phyla have a double membrane, supporting a diderm first history of bacteria, due to some loose taxonomic rules and difficulty separating the genetic diversity of these organism it is still not possible to say definitively which evolved first [29].

1.1.5 Outer membrane structure and LPS

C. jejuni is a Gram-negative bacterium characterised by a unique outer membrane. The cell envelope of Gram-negatives is comprised of two lipid bilayers; a symmetrical phospholipid inner membrane and an asymmetric outer membrane which has a phospholipid inner leaflet and an outer leaflet which contains lipopolysaccharides (LPS) or lipooligosaccharides (LOS). LPS is a charged amphipathic molecule which forms an integral part of outer membranes semi-permeable barrier – protecting the cell from a range of toxic compounds that it may encounter, including antibiotics and solvent molecules [30] [31]. The LPS structure is comprised of three components; lipid A, which is a hydrophobic moiety that anchors LPS to the outer membrane; a central oligosaccharide unit, comprised of an inner and outer core; and an extended O-antigen unit made of a repeating glycan polymer (Figure 1.1). LPS, otherwise known as endotoxin because of the strong immune response it causes in animals, is essential in most Gram-negative organisms, with exceptions including *N. meningitides* and *Acinetobacter baumannii*. Sandwiched between the two membranes is a thin layer of peptidoglycan, which is otherwise found as a thick external layer in Gram-positive bacteria. Also known as murein, peptidoglycan forms a polymer mesh from repeating sugars and amino acid units further enhancing the selective permeability and structure of the membrane system in Gram-negative.

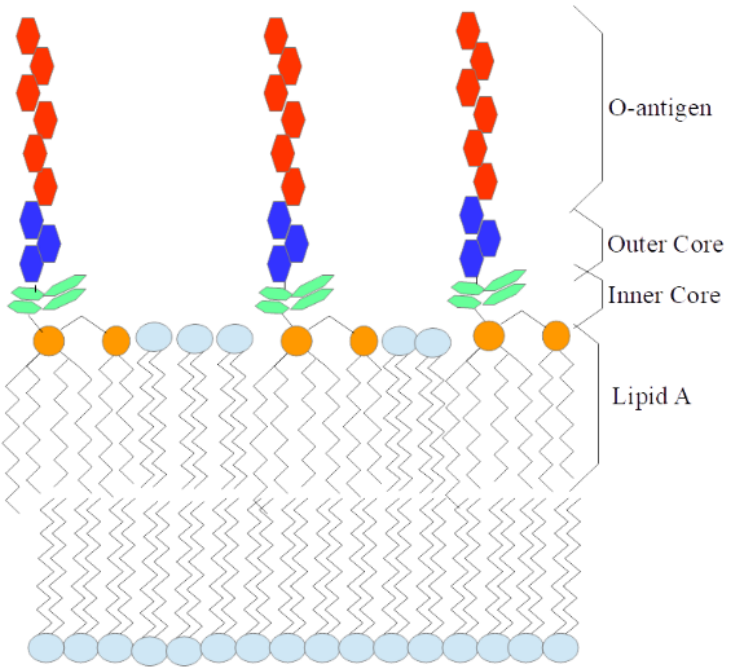


Figure 1.1 Schematic of the Gram-negative outer membrane outer leaflet. The outer leaflet of the Gram-negative outer membrane is asymmetric; the periplasmic face is composed of phospholipids, while the extracellular face is rich in lipopolysaccharide (LPS) molecules. The structure of LPS is complex, with molecular weights ranging from 10 – 20 kDa. It is shown here with lipid A, the core oligosaccharide (inner and outer), and the extended glycan O-antigen. LOS, found in *C. jejuni*, is a lower molecular weight molecule which lacks the extended O-antigen repeat shown here.

Campylobacter's distinct outer membrane is characterised by diverse polysaccharide structures. In *C. jejuni* though, a lower molecular weight LOS is found in the outer membrane. LOS lacks the O-antigen found in lipopolysaccharide, which is more common in other bacterial species, and contributes to host cell invasion in humans as described previously (Section 1.1.1) [32].

1.2 Outer membrane biogenesis

As described above, the Gram-negative outer membrane is unique as it features LPS (or the shorter LOS) and β-barrel proteins. Assembly of the outer membrane is a complex process dependant on many components but in recent years, significant work has been done towards understanding how this occurs. They are broken down into several distinct, yet overlapping, systems, which form a single network that functions to build the outer membrane. As most of the pathways are usually studied in isolation, it is unclear how all the events are sequenced and at times the presence of multiple essential genes makes this even more difficult to determine. Nevertheless, based on the dependency of some units, and essential nature of others, an arrangement for an *E. coli* model is described, starting at the inner membrane translocation units and ending with complete outer membrane biogenesis through the assembly of mature LPS (Figure 1.2).

1.2.1 Inner membrane secretion

Both precursor outer membrane proteins and periplasmic chaperones required for LPS assembly are synthesised in the cytoplasm and targeted for translocation across the inner membrane via the Sec pathway. The inner membrane Sec translocation pathway is comprised of seven unique proteins that form four units; an ATPase (SecA); a chaperone (SecB); an integral membrane complex (SecY, SecE, and SecG); and additional membrane proteins, (SecD and SecF). SecB is a cytoplasmic chaperone. It carries translated pre-proteins from the ribosome in an unfolded state to the inner membrane, which are then targeted for secretion via the Sec pre-protein translocase pathway by an N-terminal signal-peptide. SecA then interacts with SecB and drives the pre-protein through a channel formed by the SecYEG complex, in an ATP-dependant process [33].

After translocation through the inner membrane, the pre-protein signal peptide is cleaved and the mature protein is released with assistance from SecDF. Nascent OMPs then go on further to interact with periplasmic chaperones DegP, Skp and SurA which form two pathways responsible for preventing aggregation of the insoluble OMPs and potentially promoting initial folding, before insertion through the Bam system [34].

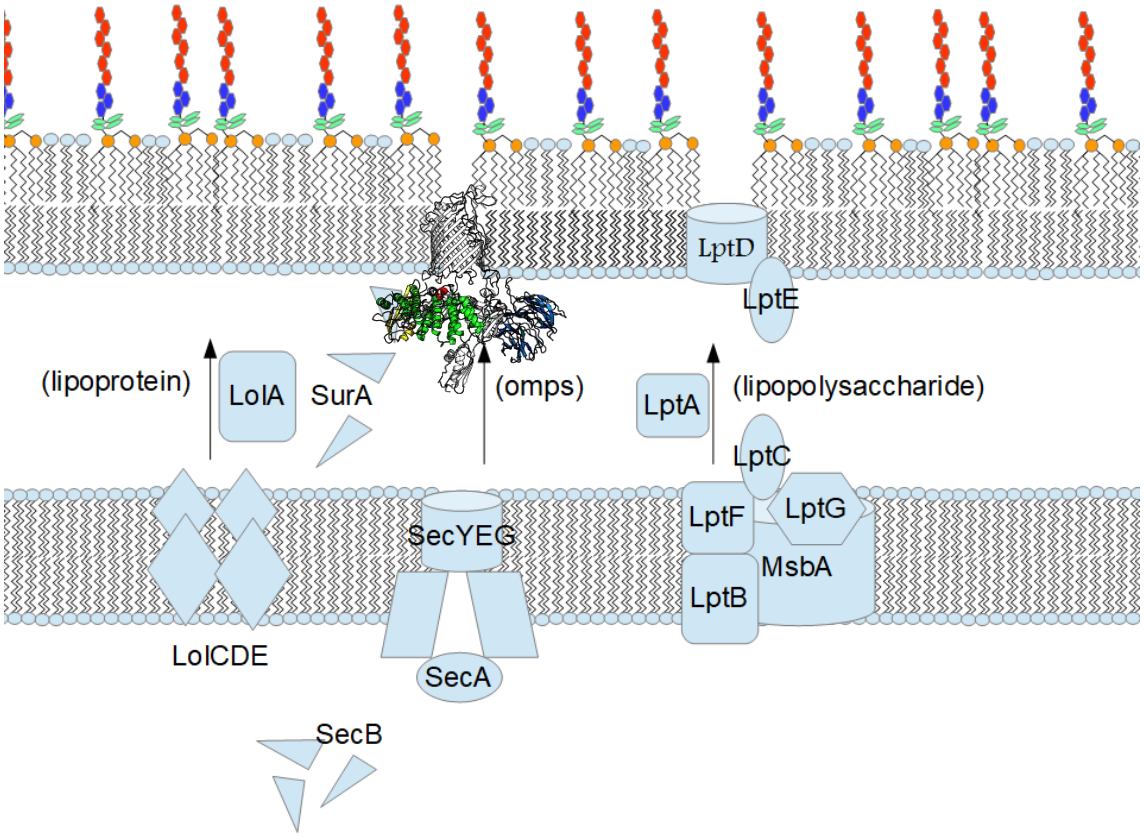


Figure 1.2 Schematic of outer membrane protein assembly in Gram-negative bacteria by the BAM complex. In the *E. coli* model, OMP synthesised in the cytoplasm first cross the inner membrane via the SecYEG. Skp and SurA assist in navigation across the periplasm before insertion in the outer membrane by the Bam system [12]. In *C. jejuni*, the homologues proteins PEB4(*Cj*0596) and *Cj*1289, equivalent to periplasmic SurA chaperones, and *Cj*0939, equivalent to pathway synonymous in most Gram-negatives studies, with homologous proteins already identified in *C. jejuni*. .

1.2.2 Lol

Bacterial lipoproteins found in the outer membrane, including components of the Bam and Lpt system, are sorted by the Lol system. The Lol system is comprised of three units; an inner membrane ATP-binding cassette (ABC) transporter (LolCDE); an outer membrane receptor, (LolB); and a periplasmic carrier (LolA). The nascent lipoproteins are expressed in the cytosol and translocated across the inner membrane via the LolCDE translocation complex. They are then recognised by LolA, which binds the lipoprotein and shuttles it across the periplasm to the inner leaflet of the outer membrane that docks with LolB. The lipoprotein is then anchored to the outer membrane via a cysteine linked phospholipid moiety.

1.2.3 Bam

Nascent OMPs are folded and inserted into the outer membrane in a passive process carried out by a highly-conserved system found in all Gram-negatives. The Bam system is comprised of five unique proteins in *E. coli,* which form the heteropentameric BAM complex; an outer membrane protein (BamA); a tetratricopeptide regulator (BamD); and three accessory lipoproteins (BamB, BamC, and BamE). The Omp85-superfamily protein BamA, which is itself an outer membrane protein, is the core hub of the integral outer membrane BAM complex. BamA and BamD, which are both essential in *E. coli,* interact to stabilise two subcomplexes formed by BamAB and BamCDE. The nascent OMP is carried across the periplasm via periplasmic chaperones [35]. Once in the outer membrane, BamA is the central component of the BAM complex, which is comprised of five known proteins BamA-E. BamA is a β-barrel protein, while BamB-E are lipoproteins that arrive through the Lol system.

Finally, once all the necessary components are in place, LPS can be inserted into the outer leaflet. LPS biosynthesis is a well characterised, conserved pathway, which occurs at multiple sites across the cell membrane, starting at the inner membrane cytoplasmic interface (see Figure 1.2). O-antigen repeats are first synthesised in the cytoplasm before being flipped across the inner membrane to the periplasmic side, attached to the lipid carrier undecaprenyl diphosphate, and catalysed by the inner membrane Wzx protein. The inner membrane WaaL ligase then catalyses the formation of mature LPS by ligation of the O-antigen repeat units to the lipid-A core, which is translocated by MsbA, at the periplasmic site where O-antigen repeats are polymerised.

The lipopolysaccharide transport system Lpt, is then responsible for LPS transport across the periplasm and outer membrane. The Lpt system is comprised of seven known transport proteins;

LptA to LtpG, which span from the cytoplasm to the cell surface [36]. The periplasmic component LptA forms a stable complex with LptC, which is anchored to the inner membrane through the LptBFGC complex. It is thought that LptC acts as a dock for LptA, which is then able to bind the mature LPS, and transported across the periplasm bound to LptA [37]. At the outer membrane, the LptA-LPS unit then interacts with LptD anchoring to the outer membrane through the homodimeric LptDE complex. Here the LPS is released from LptA and finally inserted into the outer membrane by LptDE. This completes the most fundamental aspects of outer membrane biogenesis, which relies heavily on coordination between several soluble chaperones, lipoproteins, integral inner membrane proteins and outer membrane β-barrel proteins, chiefly BamA.

1.2.4 Structure of the β-Barrel Assembly Machinery

BamA is the designated Omp85 protein and the main hub of the BAM complex in Gram-negative bacteria. Its molecular architecture is characterised by a soluble N-terminal domain, rich in polypeptide transport associated (POTRA) repeats, called the POTRA domain, which protrudes into the periplasm, and a C-terminal 16-stranded β-barrel structure, which spans the phospholipid bilayer of the outer membrane [38]. The BAM complex in *E. coli* is completed by four accessory lipoproteins, BamB, BamC, BamD and BamE, with a total size of just over 200 kDa. Although the exact roles of these auxiliary components are still uncertain, recent studies have revealed the structure of the entire complex along with interaction sites between the components, which are important details required for solving the mechanism of OMP assembly [39].

A lack of structural data for the β-barrel region of BamA had previously limited investigations of the assembly process. However, two crystal structures of β-barrel domains, one from a full-length *N. gonorrhoea* BamA with five POTRA units, and the other from a truncated *H. ducreyi* BamA with three POTRA units, have since been solved, revealing important conserved structural features [40]: both β-barrels had sixteen, anti-parallel β-strands closed by a reduced hydrophobic interaction between the first and last strands.

This reduced interaction between strands β-1 and β-16 could hypothetically function as a gate for lateral opening to facilitate the insertion of OMPs into the outer membrane. Lateral openings in outer membrane proteins have previously been proposed as a mechanism to facilitate the transport of some small hydrophobic molecules across the bacterial outer membrane [41] [42]. It was also observed in BamA that there were two conformations of the POTRA domains in relation to a concealed pore at the base of the barrel. The *Ng*BamA structure was observed in a conformation

in which the POTRA5 sub-domain interacts with periplasmic loops (pL3, pL4, pL5 and pL7) stabilised in a closed conformation, which would block access to the β-barrel pore. In the second conformation, the POTRA domains of *Hd*BamAΔ3 are rotated approximately 70° outward, interacting with pL1, in an open conformation (Figure 1.2).

These observations made from resolved BamA structures indicate that the linker region between the POTRA and β-barrel domain is flexible and could pivot to regulate access to the β-barrel. Small angle X-ray scattering and solution NMR have also shown conformational flexibility between the P2 and P3 subdomains [43]. Furthermore, it was found that the POTRA domain extended to the approximate width of the periplasmic space, which may enable it to target OMPs directly from the inner membrane after translocation [44].

Both crystal structures confirmed that the topology of *N. gonorrhoeae* and *H. ducreyi* BamA POTRA sub-units have a common β-α-α-β-β fold (Figure 1.3), as seen in other Omp85 proteins. Previous reports were made about the conserved structure and function of these units [45] [46] [47] and prior to the release of the full-length structure with the β-barrel, units from *E.coli* Omp85 YaeT POTRA domains had also been solved by X-crystallography [48] and solution state NMR [49]. The surfaces of these globular subdomains present binding sites to cargo OMPs and other auxiliary proteins in the complex after delivery by the chaperones and the shape of β-α-α-β-β seen in POTRA domains is characteristic for proteins which mediate protein-protein interactions [49].

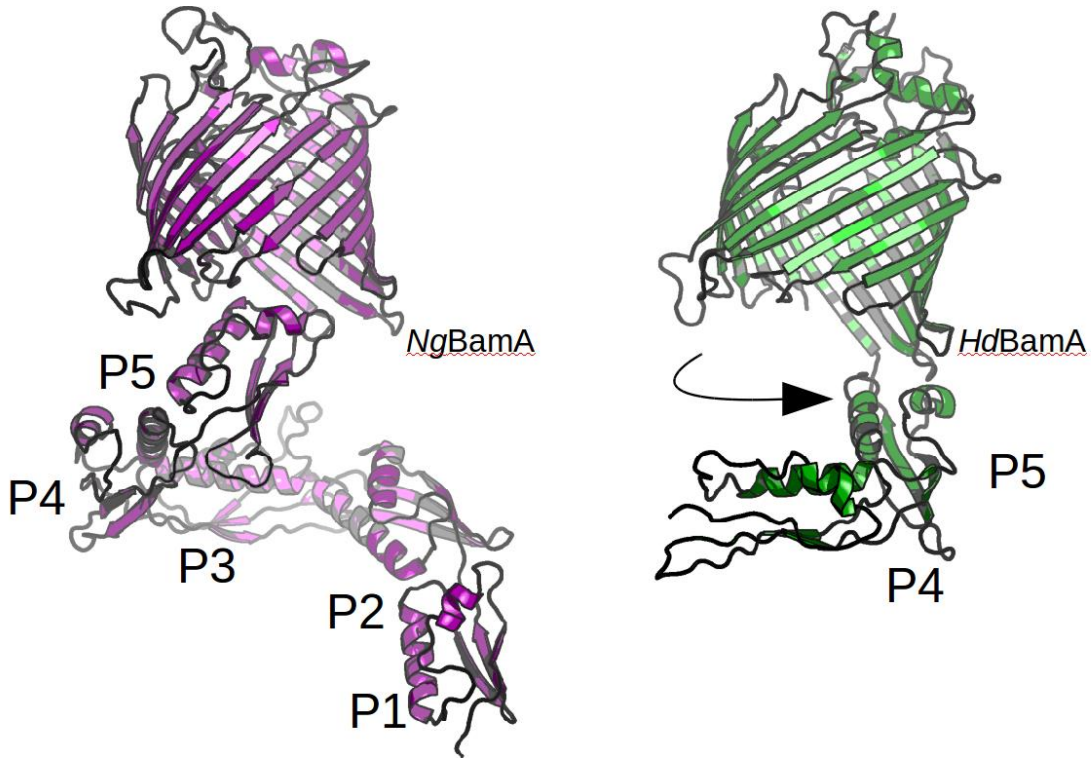


Figure 1.3 Cartoon ribbon structures of two BamA proteins solved by X-ray crystallography display unique open and closed conformations. The two conformations of the BamA POTRA domains shown here indicate that the opening of the β-barrel in a truncated *H. ducreyi* crystal is accessible in the truncated open position (right, PDB: 4k3c) whereas POTRA domains block the pore in a closed position in the *N. gonorrhoeae* structure (left, PDB: 4k3b). It is possible that the POTRA domain therefore rotates to allow OMPs access to the outer membrane and that this process could be dependent on the presence of the P123 subdomains. Visualisation and rendering was done using PyMOL.

1.2.5 Mechanism of β-Barrel Assembly

Although recent studies have revealed the structure of BamA and the BAM complex, the mechanism of β-barrel assembly is still unclear. A number of model mechanisms of OMP insertion have previously been proposed [50];

1) The pore-folding mechanism proposes that the POTRA domains and accessory components feed the nascent OMP to a pore formed by BamA, which threads the cargo into the outer membrane.
2) The complex pore-folding model, where the pore is formed by a central channel between multimeric BAM complexes that releases the cargo as the multimeric structure dissociates.
3) The barrel folding mechanism proposes that the BamA barrel structure acts as a template for β-barrels which are then translocated by accessory components.
4) The chaperone-folding mechanism suggests that the BAM complex is not involved in folding, but only insertion into the membrane, while DegP folds and protects the OMPs from degradation in the periplasm
5) The accessory folding mechanism model puts forward the idea that the nascent OMP is folded by the BAM complex before being inserted by the DegP protein or some unknown mechanism. This model suggests that the DegP pathway can then act as a quality-control step for improperly folded OMPs.

Some of these model mechanisms now appear unlikely after reviewing recent findings made in the field. For instance, the complex pore folding model (2) proposed is difficult to explain due to steric restrictions imposed by accessory BAM components seen in the complex crystal structure [51]. On the contrary, the β-barrel folding mechanism has been further supported by measurements of the pore size formed by BamA, which closely match those of a β-hairpin and supports the idea that it provides a template for nascent OMP folding.

1.2.6 β-barrel proteins (Omp85-superfamily) in other domains of life

Omp85 proteins were first discovered in the late 1990's in Neisseria strains of bacteria [52] and recognised as closely related sequences of two previously observed genes, labelled Oma87 and D-15-Ag in *Pasteurella multicida* and *Haemophilus influenzae* respectively [53]. These two genes would later be reclassified as BamA proteins and members of the Omp85 superfamily. Early sequencing results eluded to the Neisserial proteins localisation in the outer membrane from characteristic carboxyl-terminal signature sequence and amino-terminal signal peptides, typical of this class of

proteins which cross the inner membrane of the Gram-negative cell. Although variations are seen between organisms, this signature recognition site is usually composed of alternating aromatic hydrophobic residues and is one of the characteristic markers of a β-barrel outer membrane protein. Termed Omp85, after their localisation and approximate size, they were later reclassified and the Omp85 name was given to encompass an entire superfamily of proteins, including molecular machines associated in bacteria such as the two-partner secretion system (TPSS) FhaC, and translocation and assembly module (TAM) TamA in bacteria[54] [55] and Sam50 and Toc75 in Eukaryotes [56] [57].

The existence of BamA homologues in Eukaryotic organisms, where they are found in the outer membranes of sub-cellular organelles, is evidence that this mechanism is highly conserved in many diverse species. Omp85 proteins share an ancient origin, thought to predate the formation of the first complex cell, with an essential function for the first bacteria [58]. In mammals and other animal species, Omp85 proteins are also essential for assembly of mitochondrial β-barrel protein [59], whereas in plants they are found in the chloroplasts [60]. Symbiotic acquisition of a simple unicellular organism, is a current theory supported by these findings and indicating a conserved mechanism which is shared amongst many organisms [61].

1.2.7 Target substrate recognition

POTRA domains of BamA recognise a signature motif which is found at the C-terminal β-strand of most OMPs, as described previously [62]. This conserved sequence consists of an aromatic hydrophobic residue, either phenylalanine or tryptophan, at the first position of the C-terminal end and alternating hydrophobic residues at positions 3, 5, 7 and 9, back towards the N-terminus. A study of the *E. coli* PhoE OMP also found that the C-terminal aromatic residue was required for efficient folding and insertion into the membrane. A synthetic peptide that corresponded to the first twelve amino acid residues of PhoE induced the BamA pore to open in the same way as the full-length cargo protein did, supporting a model in which this region of the nascent OMP is involved in the targeting and activation mechanism.

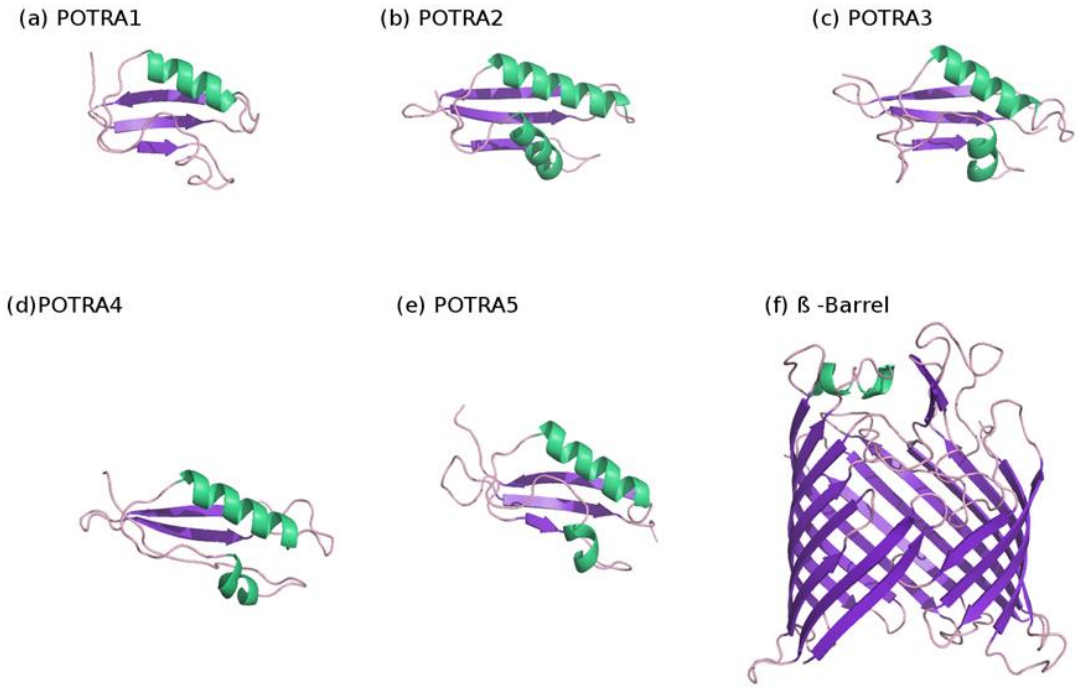


Figure 1.4 Subunits of BamA POTRA and β-barrel domain. POTRA subunits are composed of a β-α-α-β-β fold (α-helical shown in green, β-strand shown in purple) with an approximate molecular weight of 7kDa each. Typically, the BamA POTRA domain has five subunits, but sequences with up to six have been registered in the Uniprot database. .

However, although the signature sequence targeting mechanism may be conserved between species, it was also found that the non-native PorA from *N. meningitidis*, and its corresponding synthetic C-terminal signature peptide did not induce the same channel opening as in *E. coli* [62]. Upon closer investigation, the sequence of Neisserial OMPs, in comparison to *E. coli* OMPs, were found to have a positively charged amino acid residue (either arginine or lysine) at the second position, indicating a change in sequence responsible for specificity of recognition signals in different species. This is supported by further evidence that substitution in position 2 of an *E. coli* PhoE OMP, from glutamine (polar uncharged) to lysine (positively charged) prevented activation of the BamA channel, and similarly the substitution of lysine to glutamine for Neisserial PorA enhanced the assembly of the protein into *E. coli* outer membranes [62].

1.2.8 BamD and Auxiliary Lipoproteins

In *E.coli*, BamA and the auxiliary lipoprotein BamD, which anchors to the outer membrane via an N-terminal lipid and binds to the soluble N-terminal POTRA domain of BamA, are the only two essential components of the BAM complex required for nascent OMP biogenesis [63] [64]. Three other lipoproteins, BamB, BamC and BamE, which interact with the essential components, are found as part of the BAM complex in *E. coli*. While not essential for cell viability, their loss can lead to low rates of folding or misfolding of substrate OMPs and poor cell survival rates [65].

By identifying the minimal set of components required for BamA assembly, another study showed that no single specific BAM lipoprotein is required for integration into a lipid bilayer *in vitro*, although some are important catalysts for this reaction [66]. While *in vivo* experiments have shown that the BAM complex facilitates BamA integration [48], this *in vitro* study used proteoliposomes prepared using detergent dialysis methods to show that a BamCDE sub-complex was more efficient than BamA at folding a BamA substrate. This shows that a preassembled BamA is not required to assemble unfolded BamA *in vitro*. Furthermore, it was shown that BamB and BamD can individually assemble BamA, while neither were able to fold an OmpA substrate [66]. This suggests that BamA may possess some unique structural features over other OMPs, that facilitate spontaneous assembly *in vivo* and that the auxiliary lipoproteins, BamB and BamD are involved in catalysing this process.

1.2.9 The Essential Nature of BamA

It has been previously shown that a number of OMPs can spontaneously fold *in vitro* into synthetic model membranes in the absence of BamA [67]. However, it is widely accepted that BamA is required for OMP folding *in vivo.* It has now been determined that the native lipids present in the *E.coli* outer membrane impose a physical barrier on OMPs, whereas using a synthetic lipid environment might not [35]. This study investigated the kinetics of folding reactions for eight OMPs from *E. coli* (OmpA, OmpT, OmpX, OmpW, PagP, OmpLA, FadL and BamA) in native-like lipid environments. This selection of β-barrel proteins had all been shown to fold spontaneously into lipid vesicles with phosphocholine head groups, which do not naturally occur in *E. coli* - the inner leaflet of the *E. coli* outer membrane is composed mainly of phosphatidylethanolamines (PD) with phosphatidylglycerol (PG) and cardiolipin. Using lipids extracted from *E. coli*, it was observed that the efficiency of folding was extremely low when compared to the use of model membranes composed of non-native lipid.

Further experiments showed that by introducing PE lipids into the synthetic membrane consisting mainly of PC head groups the folding reaction rate was significantly reduced [35]. This effect on the folding kinetics, tested with four OMPs, was also shown to be dependent on the amount of PE introduced and a similar effect was also observed when PG lipids were introduced into PC membranes. It was therefore understood that the presence of PE and PG lipids in the *E. coli* membrane imposes a physical barrier on the spontaneous folding of OMPs, therefore necessitating the existence of the BAM complex. However, pure PC membranes did not cause this restriction and nascent OMPs could penetrate the headgroup layer.

Furthermore, this restriction on β-barrel OMPs to spontaneously fold into bacterial membranes is assumed to have evolved as a mechanism to prevent incorporation into the inner membrane that would create pores, dissipating the proton motive force and rapidly leading to cell death [35]. Molecular dynamics simulations have shown that embedded BamA structures distort a synthetic membrane around the reduced interaction between β-strands 1 and 16 (Figure 1.5), which is likely to be involved in facilitating this mechanism by destabilising the kinetic barrier of the outer membrane [68].

1.2.10 Species specificity in target recognition and folding

Other recent work includes an investigation into the species specificity of BamA in five proteobacteria [69]. Prior to this study, it was shown that BamA targets a species-specific C-

terminal signature sequence on substrate OMPs *in vitro*. This has been demonstrated *in vivo* by investigating the function of BamA homologs from *N. meningitidis, N. gonorrhoeae, B. pertussis, B. mallei,* and *E. coli* expressed in *E. coli* and *N. meningitidis* host cells. Firstly, it was suggested that an *E.coli* BamA (*Ec*BamA) is unable to substitute for *Nm*BamA of the *N. meningitidis* HB-1 strain. It is proposed that *Ec*BamA cannot properly assemble into the meningococcal outer membrane, and this was supported by testing the extractability of the foreign BamA protein from the HB-1 membrane with urea. *Ec*BamA was easily extracted from isolated *N. meningitidis* membranes, whereas this was not possible when correctly assembled in an *E. coli* membrane, indicating it had not integrated in the non-native membrane environment. Further supporting evidence was provided by investigating the susceptibility of the β-barrel domain to trypsin digestion. When expressed in HB-1, *Ec*BamA was susceptible to digestion with trypsin showing that it was not protected by a lipid membrane and, therefore, had not assembled into the outer membrane.

Similarly, it was shown that *Nm*BamA was unable to substitute for the native BamA in *E. coli* [69]. However, unlike the earlier scenario described, it was observed that a meningococcal homolog could insert in the *E. coli* outer membrane. Interestingly, when *E. coli* expressed *Nm*BamA with an N-terminal histidine tag and purified from cell envelopes by immobilised metal affinity chromatography, it was also shown that significant amounts of *E. coli* BamD had also co-purified. This supports the idea that the meningococcal BamA protein could assemble into the *E. coli* outer membrane, however it was not able to perform the role of its native counterpart, possibly due to an inability for host substrate recognition. Further experiments also showed that neither *B. pertussis* nor *B. mallei* BamA proteins could substitute for the native *Ec*BamA. However, when two closely related species were selected, *N. meningitidis* and *N. gonorrhoeae*, whose BamA homologs share 95% sequence similarity, it was shown that *Nm*BamA can be replaced by *Ng*BamA. This is an interesting observation, which shows that the evolutionary distance between species is directly related to their ability to both assemble in different environments, as well recognise nascent OMP substrates.

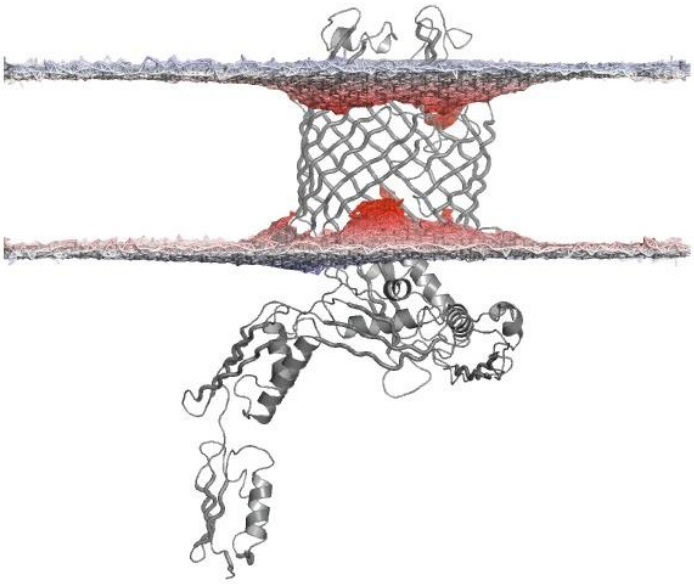


Figure 1.5 Membrane distortions observed in molecular dynamics simulations. Distortions shown in red indicate localised regions of reduced stability around the base of the β-barrel. This region between the first and last stands is critical in the putative mechanism of OMP assembly as it generates a reduced hydrophobic interaction which provides a route to the outer membrane by lateral opening. Simulation and image was created by the MemProtMD initiative (https://memprotmd.bioch.ox.ac.uk).

1.2.11 Dynamic behaviour of BamA

The use of solid-state NMR (ssNMR) has revealed new details about the mobility of the periplasmic POTRA domains of BamA proteins [70]. Magic angle spinning (MAS) ssNMR has previously been shown to be a powerful technique for studying membrane proteins in a native-like lipid bilayer environment [71] [72]. Reconstitution of uniformly ^{13}C, ^{15}N labelled BamA in pure PC liposomes, at a low lipid to protein ratio of 25:1, was performed to mimic the OM with folding and reconstitution monitored by heat modifiability using SDS-PAGE [73], susceptibility to protease digestion and extractability by urea. N-terminal POTRA domains P1 and P2 were observed as rigid features in dipolar based ssNMR experiment when associated with a lipid bilayer. Previous experiments on folded protein precipitate had shown that the POTRA domains displayed fast motion, which may be stabilised when prepared in a lipid bilayer [74].

It was also observed that the β-barrel domain of BamA displays some conformational plasticity, and electron microscopy shows that this region can disturb the lipid bilayer. A disruption of the membrane, caused by the asymmetric hydrophobic thickness of the β-barrel domain, was also shown in molecular dynamics simulations [40] and this destabilisation could be an important attribute in the mechanism by which substrate OMPs are inserted in the OM. While ssNMR experiments with phosphatidylcholine (PC) proteoliposomes are somewhat more representative of a native structure compared to crystallographic studies, the lack of auxiliary lipoprotein components of the BAM complex as well as native lipids, LPS and peptidoglycan still restricts the investigation of this mechanism *in vivo*. Furthermore, despite the wide range of studies of BamA in different organisms, no studies have yet been conducted for *C. jejuni* BamA.

1.2.12 Outer membrane protein biogenesis in *C. jejuni*

Currently very little is known about OMP assembly in *C. jejuni*. However, β-barrel OMPs are associated with essential import and export processes, and multidrug efflux functions that play a role in antibiotic drug resistance. For this reason, characterising the Bam system in this organism and other prevalent pathogens is considered to be an important area of research.

It is clear that OMPs including BamA are prevalent in the ε-proteobacteria class, as they are in all other Gram-negative classes [58]. Previously, the *Cj*0129c gene (Omp85 superfamily) was identified alongside 454 other unique membrane-associated protein from an isolated gastrointestinal *C. jejuni* strain (JHH1) using complementary proteomics technology [75]. In this

study, several other important homologs were recognised including the virulence factors PEB1A, PEB2, PEB4 and *Cj*1289.

More recently, the structure of PEB4 and *Cj*1289 have been solved by X-ray crystallography, showing that these two periplasmic protein are structurally related SurA-like chaperones and part of the analogous system found in *E. coli* which carries OMP to the outer membrane [2]. Furthermore, the Sec protein secretion system has already been characterised in *C. jejuni* [76].

1.3 Methodology

While it is understood that OMP biogenesis should be essential for survival, only BamA was identified previously through sequence analysis. The overall aim of this study is therefore to better characterise the system in *C. jejuni* NCTC-11168, a model *Campylobacter* organism, and identify missing components of the BAM complex in the Parkhill genome [20].

The central hypothesis is that, as a Gram-negative bacterium, in order to survive, *C. jejuni* must possess a system analogous to the β-barrel assembly machinery complex found in *E. coli*. Furthermore, based on the assumption that the mechanism of function is conserved, any homologous BAM proteins found in *C. jejuni* should share structural and functional properties seen in the crystallography data from published literature. As membrane protein structure and function are challenging to determine though, a comprehensive experimental approach is proposed to learn about the structure of any *C. jejuni* Bam genes identified.

Previously methods were developed by D. Griffin (University of Nottingham) in the lab to overexpress and purify *C. jejuni* BamA from inclusion bodies using strong chemical denaturants including urea and guanidine hydrochloride, adapted from published materials and methods used to successfully express, purify and crystallise the 12-strand β-barrel NalP that is part of a *N. meningitidis* autotransporter virulence mechanism [77]. However, although the expression system produced high yields of relatively pure protein, downstream processing stages were not optimised for our target structure and the presence of large protein aggregates was not eliminated. For this reason, atomic force microscopy, dynamic light scattering, and SDS-PAGE will also be used to investigate the oligomeric state of BamA after purification using these methods. Following this, a new process is proposed encouraged by methods used to crystallise and the release of the first full-length crystal structure of BamA (PDB code: 4K3B) [40]. It was determined that targeting the

protein to the outer membrane by utilising the native *E. coli* translocation pathway by incorporating a pelB leader sequence could increase the chance of obtaining high quality structural data by extracting the target protein, folded in the outer membrane. The pET20b plasmid that is used for expression of *Ng*BamA and *Hd*BamA was kindly provided for by the S. Buchannan group (NHI, Baltimore, USA).

1.3.1 Aims and objectives

A combined *in vitro* and *in silico* approach is proposed to investigating structure and mechanism. The availability of high performance computing facilities also warrants the use of molecular dynamics programmes for studying the mechanistic behaviour in a model system. Online database searches can determine the existence of accessory BAM lipoprotein components in *C. jejuni* which, prior to commencement of this work, were unidentified. The outcome of these searches could be used to further model and study protein-protein interactions with cargo and within the BAM complex.

Experimental

1) In order to carry out experimental work and resolve a high-resolution structure, the sequence of *C. jejuni* NCTC-11168 BamA needs to be identified and cloned into a suitable vector for expression. The approach to this, along with methods for optimisation of expression, isolation, and purification, to obtain suitable quantities of purified protein for biophysical characterisation, is detailed in Chapter 2 and Chapter 3.

2) To study the behaviour of the protein in a native-like environment, reconstitution of purified BamA protein in synthetic lipid bilayers needs to be performed. This process needs to be monitored by assessing particle size distribution using dynamic light scattering and atomic force microscopy to determine oligomeric state, the results of which are shown in Chapter 3.

3) To assess the structure and folding of the purified protein, sample solutions also need to be studied using suitable biophysical techniques. The background, methods and results used to determine this, along with methods to determine the stability of the protein in different detergent micelle conditions are shown in Chapter 5.

Computational

1) In order to support and validate the results from experimental findings, template-based homology models for *Cj*BamA should be generated using the different protein modelling software tools available. The methods and results of this, and other methods used to study the structural relationships between BamA and other the Omp85 superfamily members are shown in Chapter 4.

2) A thorough bioinformatics search, using sequence alignment tools needs to be carried out to investigate the potential existence of homologous BamBCDE components in *C. jejuni*, which were previously unidentified. The methods used to identify a homologous BamD protein and the results of models generated using this sequence are also found in Chapter 4.

3) In order to test the structural integrity of BamA and BamD models, as well as investigate the behaviour of previously solved X-ray crystal structures of BamA proteins, molecular dynamics simulations will be used to study the essential dynamics and functional properties of these models. The background, methods and results of simulations, and how these may be used to further our understanding of OMP assembly in Gram-negative bacteria is presented in Chapter 6.

1.3.2 Research techniques

An array of biophysical techniques are currently available to study the structure and function of proteins, including;

- X-Ray Crystallography Diffraction (XRCD)
- Solution Nuclear Magnetic Resonance (NMR)
- Solid-state Nuclear Magnetic Resonance (ssNMR)
- Cryo-Electron Microscopy (cryo-EM)
- Circular Dichroism (CD)
- Atomic Force Microscopy (AFM)

As the cloning, expression and purification steps first need to be addressed, the remaining experimental work should be focused on resolving the structure and further investigating the

mechanism of membrane protein assembly, using a select number of biophysical techniques that are appropriate for a heterogeneous membrane protein. Both XRCD and ssNMR are suitable for this and have been used to investigate BamA and other Omp85 proteins, as shown previously [40] [70] [78]. Furthermore, some techniques, namely NMR, atomic force microscopy and circular dichroism can be used to study dynamic changes in structure and function. However, these methods rely on high-end instrumentation for which access can be a limiting factor. For this reason, after Experimental Aims (1) and (2) are completed, the folding and structure will be investigated using XRCD and circular dichroism. The results of these experiments will be analysed alongside models from readily available bioinformatics tools, which are also presented in this study along with the lab-based experimental methods. A pre-existing collection of sequenced genomes for *C. jejuni* and other closely related *Campylobacter* species has already been registered in the public domain, providing sufficient genetic information to exploit modern bioinformatics approaches to make this possible. These methods include protein sequence alignments, co-evolutionary signal detection and protein structure modelling tools that are adopted throughout and used to support experimental findings.

Chapter 2. Materials and Methods

2.1.1 Buffer preparation

Buffers were prepared using a 10X stock solution of TBS for 50 mM and 200 mM Tris- working concentration using reverse osmosis MilliQ water and the pH adjusted to 7.5 with hydrochloric acid. Solutions were filter sterilised using a 0.20 μm filtration unit before use and stored at 4°C for up to 4 weeks.

2.1.2 Batch culture preparation and inoculation

Terrific-Broth (TB) nutrient mix (24 g/L yeast extract and 12 g/L tryptone) was prepared in well rinsed, clean Erlenmeyer flasks up to 900 mL and autoclaved at 121°C for 12 mins before being left to cool to room temperature, followed by the addition of 10X TB salt solution (100 mL, 0.17 M KH_2PO_4, 0.72 M K_2HPO_4) and 1000X sterile stock ampicillin solution (1 mL, 100 mg/mL). Overnight growths were prepared as described previously (Section 2.1.4) and 1 mL was used to inoculate each 1 L flask of TB media.

2.1.3 Low temperature induction

Cells were grown for up to 4 hrs at 27°C, 200 rpm until an OD_{600} (optical density absorbance measurement, 600 nm) of around 0.4 was reached before the temperature was reduced to 16°C. After 1 hr, to allow for the liquid temperature to equilibrate, protein expression was induced using 0.4 mM (final concentration) filter sterilised IPTG from a sterile 1 M stock, stored at -20°C.

2.1.4 Harvesting

Cells were harvested in a J-25 series floor standing centrifuge (Beckman Avanti) at 4,500 g for 30 min at 4°C. The supernatant was discarded, and pellets were rinsed gently with pre-chilled lysis buffer (20 mL) to remove excess media. The cells were then weighed and resuspended in 5X lysis buffer (5 $mL/g_{wet\ cell\ weight}$) using a repeated gentle pipetting action.

2.1.5 Lysis

Resuspended cells were kept on ice and lysed with three passes through a pre-chilled French press at 20,000 psi (~137,900 kPa). Alternatively, cells were lysed using one pass through a Constant Cell Disruptor (Constant Systems, UK) with a pressure setting of 36,000 psi (~248,210 kPa).

2.1.6 Outer Membrane Isolation

Cell lysate was split into 50 mL Falcon tubes and cleared at 12,000 g for 10 min at 4°C in a temperature-controlled Avanti benchtop to remove undisrupted cells and debris. The cleared supernatant was then recovered and incubated with 5% Elugent (Calbiochem, USA) in a glass dish at room temperature for 30 min with gentle mixing using a magnetic plate stirrer. Ultracentrifugation at 180,000 g for 1 hrs 30 min was applied in 12 ml polycarbonate centrifuge tubes using a SWTi-41 rotor (Beckman Coulter) or A840 rotor (Sorvall) in a floor-standing ultracentrifuge. The supernatant, containing solubilised inner membrane fractions, was discarded and the pellet containing purified outer membrane fraction was resuspended in TBS solution with detergents (Section 2.2.7). The sample was then spun at 200,000 g for 1 hr to remove insoluble fragments before loading for FPLC (Section 2.2.8).

2.1.7 Crude Membrane Isolation

Cell lysate was cleared as described above and ultracentrifugation at 160,000 g, 4°C for 2 hrs was used to pellet total membrane fragments. At this point the quality of the cell lysis method was assessed by observing the composition of the pellet. The pellet was solubilised in 2% (w/v) DDM in TBS buffer by hand using a combination of a P1000 to gently disrupt the large pellet, followed by multiple passes through a thin gauge needle until the solution was free from large clumps. The solution was then spun at 200,000 g to remove insoluble matter.

2.1.8 Immobilised Metal Affinity Chromatography

AKTA Prime Plus System

Affinity column chromatography was performed on an AKTA Prime Plus (Pharmacia, Sweden) using prepacked Superflow Nickel-NTA columns (Qiagen, Holland). Before each trial, the system was first washed with 20 % EtOH and flushed with MilliQ water. The system was equilibrated in TBS with 0.1% n-Dodecyl-β-D-maltoside (DDM, Sigma-Aldrich, UK) buffer and 20 mM imidazole to reduce non-specific binding. The sample was then loaded at a flow rate of 0.1 ml/min to maximise residence time in the column. Once the sample was loaded, the column was washed with 20-30 column volumes of wash buffer, TBS with 0.1% DDM and imidazole (20 mM). BamA

protein was eluted using a linear gradient from 0 – 250 mM imidazole over 10 mL while 1 mL fractions were collected in clean, glass test tubes.

Loose Resin Purification

IMAC was performed using loose Ni-IDA resin (BioRad, USA) in 15 ml Falcon tubes at 0.05% (v/v). The resin was first washed with MilliQ water to remove 20% EtOH storage solution. 5x times the equivalent resin volume was then added and mixed by gentle rocking for 2 mins. The resin was pelleted by centrifugation at 1000 g for 30 seconds and the supernatant stored for SDS-PAGE analysis. This washing procedure was repeated a further two times. Binding buffer, TBS with 0.1% DDM and 20mM imidazole, was added to 5x the equivalent resin volume and the resin was then left to equilibrate for a further 2 mins. The resin was pelleted again using centrifugation at 1000 g for 30secs and the supernatant discarded. The crude detergent solubilised membrane solution, containing BamA, was then incubated with Ni-IDA resin for 16 hrs at 4°C using gentle rocking. After incubation, the resin was pelleted, and supernatant collected for SDS-PAGE analysis. BamA protein was eluted with imidazole TBS buffer (250 mM) in a 2 ml PD10 Biospin column (BioRad) manually using a P1000 micropipette. Fractions (1 mL) were collected and kept at 4°C.

2.1.9 Size Exclusion Gel Permeation Chromatography

SEC-GPC was performed using a prepacked Superdex-200 column on an AKTA Prime (GE Health-care) at the Membrane Protein Lab, Diamond Light Source, Oxfordshire. The system was first washed and equilibrated with 0.03% DDM in TBS solution. The sample was then loaded at 1 mL/min and fractions were collected in 1 mL vials.

2.1.10 Protein Concentration Determination

Elution fraction protein concentrations were determined using a Nanodrop-2000 (Thermo-Fischer, USA) blanked with elution buffer. The measurements were used to plot an expected elution curve for fractions collected from manual IMAC procedures to determine when the protein had eluted completely. Peak elution fractions were collected and buffer exchange in a 100 kDa MWCO ultrafiltration device was used to remove excess imidazole. The protein concentration was then assessed using the same method for the pooled peak fractions, blanked with TBS containing 0.1% DDM.

2.1.11 Liposomal Reconstitution

Protein samples prepared in DDM detergent micelles were added to a solution of DMPC liposomes at a 10:1 molar ratio of lipid to protein. Lyophilised 1,2-dimyristoyl-sn-glycero-3-phosphocholine (10 mg, DMPC, Avanti Polar Lipids, USA) powder was resuspended in chloroform (1 mL) in a glass round-bottom flask. Rotary evaporation was then used to remove the excess liquid solvent. The resulting lipid cake formed was then hydrated in TBS buffer (1 mL). BamA protein, concentrated using a 100 kDa MWCO Vivaspin ultrafiltration device (Sartorius, Germany) was added to the liposome solution and dialysed against TBS for 72 hrs at room temperature, replacing the dialysis buffer twice daily.

Chapter 3. Membrane protein expression, purification, and characterisation

3.1 Introduction

The main role of the bacterial membrane is to encapsulate cellular components, while forming a protective barrier to defend against environmental changes. However, in order to facilitate essential transport functions, lipid bilayers must be semi-permeable to permit the flow of nutrients, ions and other essential small molecules into the cell while also allowing waste products and toxins to leave the cell. In most cases, embedded membrane proteins are needed to achieve this. Transmembrane proteins are divided into two distinct classes; α-helical bundles which are found spanning the inner membrane, and β-barrels which are found almost exclusively as outer membrane proteins [7]. While α-helical transmembrane protein can fold into their native environment spontaneously, β-barrels are unable to do this and a reliant on the Bam system for assembly

To understand the mechanistic behaviour of the Bam system in an organism, it is first necessary to learn about its structure. In fact, the idea that '*structure is function*' is a popular maxim amongst structural biologist [79]. Although solving the three-dimensional structure of a membrane protein is a notoriously difficult task, gathering structural information for these biological macromolecules remains an important step towards understanding membrane protein activity and is valuable for various areas of research with applications in medicine, industrial chemistry and biotechnology [80]. In this section, methods of expression, purification, and reconstitution of *Cj*BamA are developed and used to obtain high purity, concentrated membrane protein samples in detergent buffer solutions, for use in structural investigation studies including X-ray crystallography. Furthermore, genetic engineering tools are applied to optimise preparative steps in a recombinant *E. coli* BL21 expression system which targets the protein for extraction from the host cell membrane. The suitability of different detergent buffer solutions is also investigated, and the relative advantages of these methods are discussed against over expression and refolding protocols which are otherwise used in membrane protein preparation.

3.1.1 Identification of BamA in *C. jejuni* NCTC-11168

The *C. jejuni* NCTC-11168 BamA target gene was first identified in the online UniProt database; Outer membrane protein assembly factor BamA, Entry: Q0PC08, 2006 using the online UniprotKB search engine (www.uniprot.org) with query terms *'jejuni'* and *'omp85'*. The sequence is 739 residues long and has a predicted mass of 83,198 Da. The first 17 residues correspond to the inner membrane translocation signal peptide, positioned at base pairs 129,800 to 132,019 on the genome. The full length, mature *Cj*BamA sequence is therefore 722 residues long and has a predicted mass of 81,401 Da. The BamA gene is found adjacent to an inositol monophosphatase at the C-terminal side and a prephenate dehydrogenase on the N-terminal side (KEGG). The sequence of the *C. jejuni* BamA protein found is as follows:

C08|Q0PC08_CAMJE Outer membrane protein assembly factor BamA OS=Campylobacter jejuni subsp. jejuni serotype O:2 (strain ATCC 700819 / NCTC 11168) OX=192222 GN=bamA PE=3 SV=1

MKKHLISICALVAMANAATIKDIKFIGLNHLSNTSAINIAGLKIGEEINPAKINTAILNLYKQNY
FENIAVENNNGILEIIVTEKPTIAKVTITGIASNDRKQVESILGIKRGTLLDEGNIKEAIERIKAY
YEAKSYFDTIVEYKKKTLENTDGLELEFIVNRGENIIIDNVHLSGAKKFSYSDIEPAVVNKEKE
FMGWMWGRNDGKLKVFELSNDSSRIADEYMKKGYLDVQVSSPYLKTYTDTYQANLTYFIKE
GKPYKIKSISIENPLFDDKQNAQTVKDLRSSAGKTINIEDIRKDVKTIETQSADLGYAFVEVYP
DIQKNDQTQEATVVFKVIPHDKVYIRNVIISGNSRTVDRVIRRELYITEGNLYNRTDLSESKNA
LKRTSYFDDVNIKEEKVDDTHIDLIVDVKEASTGAISGGIGYGSSDGILLNASLSDTNIFGSGIK
SSVSVDKSDDTLSGRISLVNPRVLDSQYSLGGTLYSNDYEWDNYSEKNYGFDITIGRQFARYY
NVSLTYNLEQSDIYHLSPTLLRTGYELGKSIKSSITPAITFNDTDDYYLPRSGIIASTSLEYAGLG
GDQEFISSSSKFNFYQGLQDYIGYDLIYRYKASFYKVWDEGYLPINQRIYLGGIRSIRGFESRTV
SPKNQWGDEIGGTIAFANSVELSFPLIDRIKLRGSVFFDYGMIGRKNLDEIKRMSTGIGIEWITP
IGPLQLVFAKPLNDKKGDDTNSFEFNLGTRF

This sequence, minus the signal peptide, is the target for structural studies and used later to generate protein homology models. Initially no other entries for auxiliary BAM components in *C. jejuni* NCTC-11168 or other Campylobacter strain, were found in the UniProt database and it was believed that BamD could be absent in ε-proteobacteria, as previously stated in literature [81]. However, it has now been confirmed that the auxiliary BamD protein is present in a number of ε-proteobacteria, for which the identification and sequence of this (*Cj*1074c) in this *C. jejuni* are presented later (Chapter 4).

3.2 Background and Theory

Membrane proteins are an important group of biological macromolecules with essential functions associated with molecular transport and signalling processes in both cells and sub-cellular organelles. To accurately determine the mechanisms by which they function, it is necessary to obtain a high-resolution view of the macromolecular structures involved, ideally at an atomistic

level and historically, X-ray crystallography has been used as the principal tool used to achieve this. Myoglobin was the first protein structure determined at high-resolution using this method [82] - a considerable achievement for which the 1962 Nobel Prize in Chemistry was awarded only a few years later - and XRCD remains one of the most reliable and popular methods for mapping the atomic co-ordinates of proteins.

However, despite significant developments in this area, protein crystallisation is still problematic and a tricky, but essential, step to complete before a crystal structure can be solved. In fact, the primary bottleneck for obtaining a structure is first associated with the challenges for growing diffracting crystals [83] - ordered arrays of packed proteins that are isolated in a dehydrated state, produced from a supersaturated, high purity and high concentration protein solutions. This is a technically challenging operation and only made more difficult when working with membrane proteins due to their largely insoluble nature. On top of this, membrane proteins are also more difficult to express and purify than their soluble counterparts, due to the need for solubilising membrane mimetics such as detergents, which are described in more detail later in this chapter (Sections 3.2.2 onwards). Furthermore, there is no guarantee that crystallisation will be successful even once suitable quality protein samples are obtained, meaning that extensive optimisation processes are usually required. For these reasons, the process scale requirements for attaining sufficient quantities of membrane proteins are also therefore usually much more demanding.

Nonetheless, once an established protocol has been developed, the methods used to express, purify, crystallise, and finally solve a protein structure tend to be reliable and reproducible. New target proteins which have structural similarities i.e. properties which may now be predicted using bioinformatics tools (Section 4.1), can also be more easily and readily attained using established methods for similar targets. In this manner, the field is moving forward quicker than ever with new structures of homologous proteins from unique organisms, as well as novel, bigger and more complex protein assembly structures released each year.

3.2.1 X-Ray Crystallography

As previously stated, X-ray crystallography is a widely-used technique to determine the atomic and molecular structure of crystals by using an intense beam of high-energy light, which is diffracted in many specific directions by a regular and repeating atomic structure. X-rays range in wavelengths between 0.01 and 10 nm, corresponding to frequencies of light from 30 PetaHz to 30 ExaHz and

energies up to 100 keV. These wavelengths are a requirement for achieving a sufficient diffraction limit needed to observe atomic displacements at Angstrom resolution.

Once a successful crystallisation process has been achieved, the resulting structure is studied using X-ray crystallography diffraction (XRCD). This method works on the principle that the atoms found of a molecule trapped in a crystalline state, and therefore restricted from diffusive motion, will diffract an incident beam of X-rays at a specific angle (Figure 3.1). Constructive and deconstructive interference occurs between the atoms in each unit cell which are used to generate an electron density map within which atom positions are placed, according to Braggs Law (Equation 3.1), to solve the overall structure:

$$n\lambda = 2d\ sin\theta \qquad (3.1)$$

XRCD can be applied to solve the structure of many kinds of materials, but it is the favoured technique for determining the structure of large biomolecules, such as DNA and proteins, at high resolution. It is considered a high-end structural investigation technique as the quality of a diffraction pattern relies on the availability of an intense, focussed beam of X-rays which is not always readily available. Nowadays, synchrotron radiation light generated at a central circular particle accelerator facility, more commonly known as a synchrotron, is used to provide more powerful X-rays for diffraction studies. Facility access is generally in high demand though and furthermore, synchrotrons are costly to construct, operate and maintain. Because of this, optimisation of membrane protein crystallisation at every stage is desirable to reduce initial processing costs and yield better crystals, increasing the chance of successful structural determination, even before beamtime is assigned. Prior to SR-XRDC crystals may be tested at smaller in-house X-ray facilities, however these are not as powerful and generally can only be used for preliminary studies on larger crystals.

For proteins, because crystals are imperfect and subject to artefacts, the resulting molecular structure is not always a good representation of how it may fold in nature. Furthermore, poor resolution can limit the accuracy of the resolved structure and information can be lost during the transformation from the scattering data to the 3D co-ordinates. For determination of the molecular structure to an atomic resolution, it is however necessary to distinguish the position of each atom from one another. To achieve this, the molecules must be held in a fixed position during imaging and, secondly, the molecules must be crystallised in the same conformation within the

crystal. However, the quality of a protein crystal will indefinitely vary with imperfections which can lead to poor resolution [84].

A value known as the R-factor is used to describe reliability or how well the diffraction data fits the resulting model structure. Additionally, a temperature factor, otherwise known as the B-factor or Debye-Waller factor, is used to describe the attenuation of scattering caused by thermal motion which causes displacement of atomic positions from the average value. As such, flexible regions in a protein will have higher B-factors and well-ordered regions of the structure will have lower B-factors. These values are often contained in the PDB files used to record protein structures and can be represented visually using computer software such as the Python Molecular Viewer [85].

3.2.2 Membrane protein extraction with detergent

Prior to and during crystallisation, detergents and other membrane mimetics are applied at various stages to solubilise membrane proteins and restrict aggregate formation. As shown earlier, membrane proteins are naturally found embedded in lipid bilayers which makes them difficult to handle outside of their native environment. Nonetheless, it can be achieved *in vitro* with the use of membrane mimetics such as detergents and synthetic lipids.

The arrangement of amino acids of a natively folded membrane protein exposes hydrophobic residue side-chains to the hydrocarbon tails of the bilayer lipids, which orientate towards each other, away from the solvent water molecules. For this reason, because of the exposure of hydrophobic regions, membrane proteins tend to be highly insoluble in water and require assistance in transport processes *in vivo* by chaperone proteins. Molecular chaperones are a diverse family of multidomain proteins which assist nascent membrane proteins in reaching their native folded state while preventing aggregation by encapsulating them to mask hydrophobic regions [86]. Similarly, detergents are water-soluble amphipathic surfactant molecules employed to solubilise cell membrane bilayers and extract membrane proteins in a similar manner, working to maintain the native structure and supress aggregation by masking hydrophobic sites.

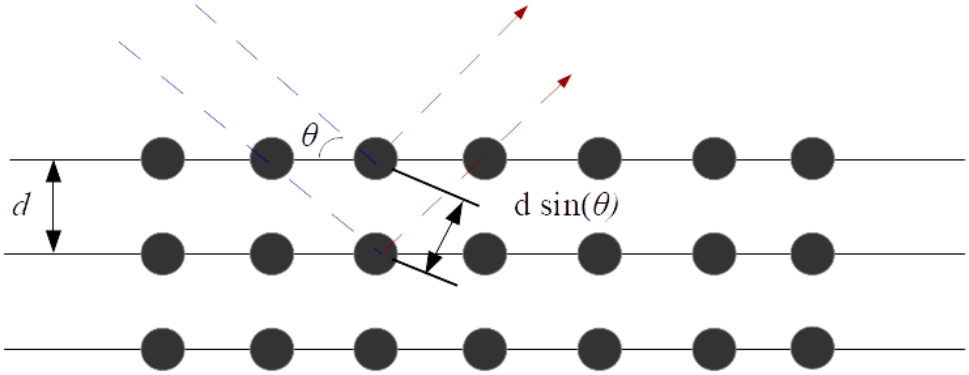


Figure 3.1 Schematic theory of X-ray diffraction crystallography. Atoms arranged symmetrically in a cubic crystal lattice, diffract the incoming light to create a reflection spot when the path length difference is equal to a multiple of the wavelength.

Detergents protect against aggregation when present at low concentrations by absorbing onto a surface or interface within a system, significantly altering the free energy in formation of the surface or interface. Surfactants typically reduce the *surface tension* of a liquid - a measure of the work required per unit area to create the boundary between a liquid and gas – or the *interfacial free energy* required between two liquids or solid and liquid in a biphasic system [1]. The free energy required to create the surface boundary is also used as a measure of the difference the two phases. The greater the dissimilarity in the chemical nature of the different phases, the greater the interfacial or surface tension between them. When the interfacial area is expanded, the minimum work required to create the additional amount of the interface therefore increases. A surfactant is, therefore, a molecule which absorbs at low concentrations at regions at or along the entire boundary area between different phases, altering the amount of work done by a significant amount. Usually a surfactant lowers the free energy, and therefore reduces the amount of work done per unit area to form the interface:

$$W_{min} = \gamma * \Delta_{interfacial\ area} \qquad (3.2)$$

where W is the amount of work done and γ is the interfactial tension. Hence, the interfacial tension can be reduced with a low concentration of surfactant while the interfacial surface area is effectively kept the same.

The physical, chemical, and electrical properties of substances are considerably different at the boundaries between different phases, therefore the interfacial free energy and surface tension are also different to those properties in the bulk phase. This means that detergents are highly effective at solubilising membrane proteins, even at low concentrations, as only a small fraction of detergent molecules make up the boundary while the rest of the population is found in the bulk of the solvent. Even so, at higher protein concentrations an increase in detergent concentration may be warranted to sustain the solubilising effect. As discussed later (Section 3.3), excess detergent can also be detrimental, interfering in the crystallisation process, as well as being a significant process cost in larger scale batch cultures up to and over 10 L.

3.2.3 Application of membrane mimetics

In addition to membrane extraction processes, detergents can also be applied to solubilise membrane protein aggregates after overexpression or for *in vitro* refolding [87]. Isolation, refolding, purification, and storage of membrane proteins can be facilitated using any number of different

commercially available detergents or detergent mixes. However, as the conditions within the bulk phase directly affect the properties at the interface, the surfactant should ideally be tailored to suit the chemical and physical properties of the system i.e. solvent composition, temperature, pH and pressure.

Detergents are amphipathic molecules with negative and positive electrical charges associated with the polar headgroups and neutral, non-polar hydrocarbon tails (Figure 3.2). They are classified into several different groups based on the properties of the charge found in the headgroup which can be ionic, non-ionic or Zwitter-ionic. Ionic detergents are either anionic or cationic, which have negatively and positively charged headgroups, respectively. Anionic detergents such as SDS and *N*-lauryl sarcosine are the most important amongst synthetic industrially manufactured detergents and commonly used in membrane protein preparation. They have a molecular structure with a long (C_nH_{2n+1}) alkyl- chain attached to an ($-SO_3$) sulpho- group forming sulphates and sulphonate salts.

Anionic detergents like SDS are found in most domestic hygiene products for their ability to remove dirt and break up bacterial cell membranes, which also makes them useful in membrane protein extraction. Because of the strong charge properties of ionic detergents though, they can also lead to protein denaturation. Non-ionic (DDM) detergents which have no net formal charge and Zwitter-ionic detergents (SB12, CHAPS), which have both anionic and cationic charges, can all be applied for solubilising membrane proteins.

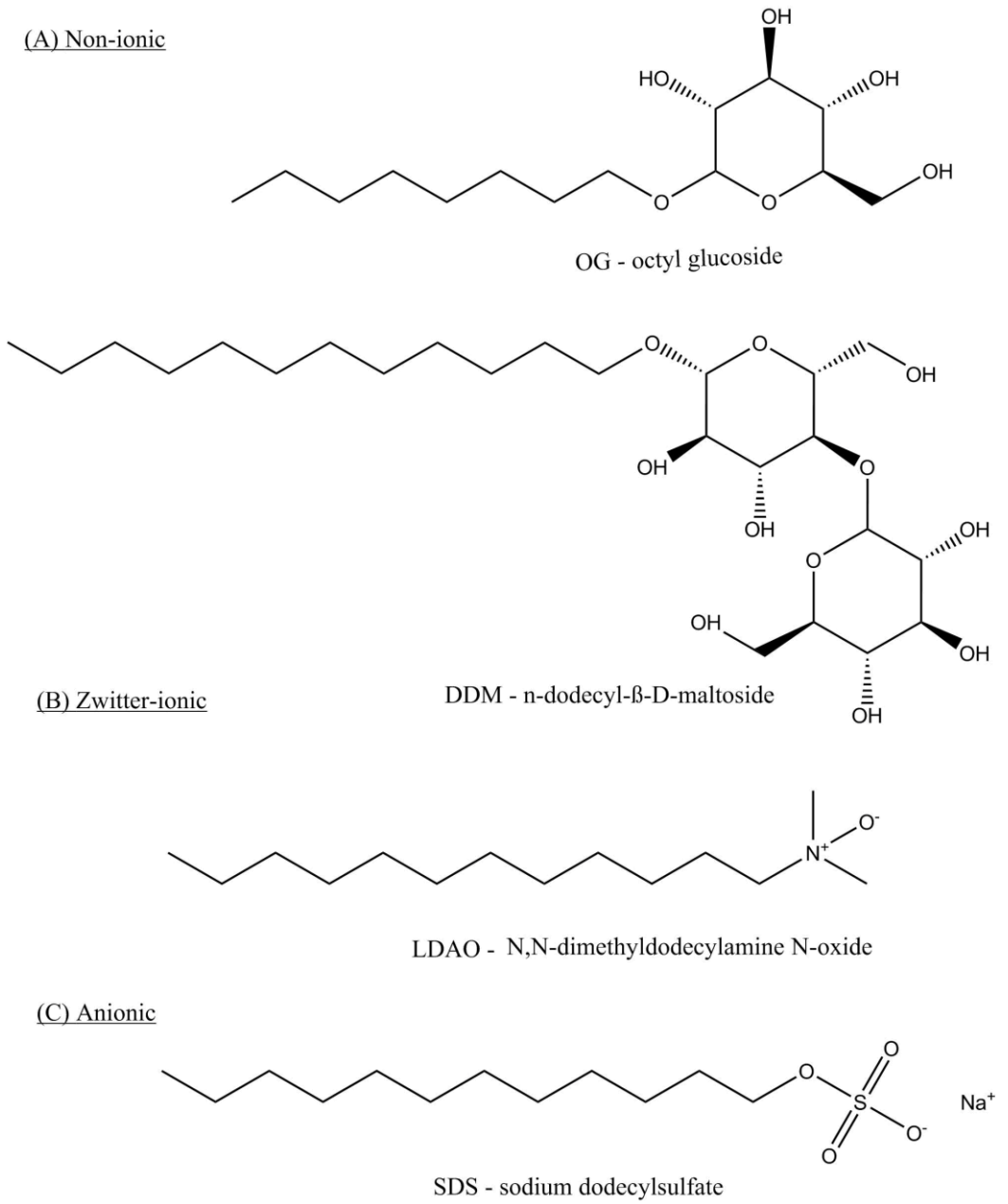


Figure 3.2. The chemical structure of detergent molecules is characterised by an extended alkane chain and a polar headgroup. The structures of detergents can be polar non-ionic, like OG and DDM (A), Zwitter-ionic, like LDAO (B) or charged with ionic headgroups, such as SDS (C), and have a hydrophobic hydrocarbon tail. This amphipathic arrangement of charges leads the molecules to self-assemble when dispersed in a solvent such as water.

3.2.4 Thermodynamics of membrane protein solubilisation

Due to their ability to fold in different ways, proteins can assume different states in solution. However, aggregated protein forms a dense, incompressible material with solid properties [88]. Detergents are typically used to reduce the amount of work required to solubilise aggregated proteins formed as inclusion bodies by coating the surface as they are reduced into smaller

particles. Furthermore, detergents then prevent aggregation events and facilitate solubilisation of membrane proteins *in vitro* by reducing the surface tension between the exposed hydrophobic residues and a buffered, ionic aqueous solution.

In solution, detergents spontaneously form ordered assemblies know as micelles. The micellar phase describes a structured aggregate of surfactant molecules which forms in the right conditions [2]. Micelles spontaneously self-assemble when the concentration of surfactant in solution is above the *critical micelle concentration.*

In solution, micelles form as a dispersed liquid colloid of aggregated surfactant molecules when the concentration is above the CMC. The free energy of surface formation in the solution is reduced by exposing the *hydrophilic* polarised headgroup to the solvent phase, while the *hydrophobic* hydrocarbon tails are sequestered from the solvent by orienting towards one another. The micellar phase is a form of soft matter arising from the packing behaviour of single chain lipids in a bilayer [89], that is the energy required to simultaneously fill the internal volume of a bilayer while maintaining a planar surface with the same area per lipid headgroup. This forces a bilayer with single chain lipids which have a conical geometry to curve. Alternative shapes of lipids can give rise to other packing geometries and phases driven through the hydrophobic effect. The force acting on the headgroups, imposed by the polar solvent phase, similarly forces detergents into micelles, known as the normal-micellar phase. The critical micelle concentration (CMC) is dependent on both the chemical nature of the detergents and the properties of the solvent in which it is dissolved. As stated previously, the boundaries between two phases is formed by a relatively small fraction of molecules compared to those in the bulk of each phase, therefore the CMC is normally quite low – typically less than 1 % (w/v) at normal room temperature and pressure. At lower concentrations below the CMC, monomeric detergent molecules can diffuse between dissociated states at a rate dependant on the free energy in the bulk phase of the solvent solution. The CMC of any given detergent is therefore affected by the chemical properties of the detergent molecule as well as physical and chemical properties of the solvent solution, such as temperature, pressure, pH, and ionic strength.

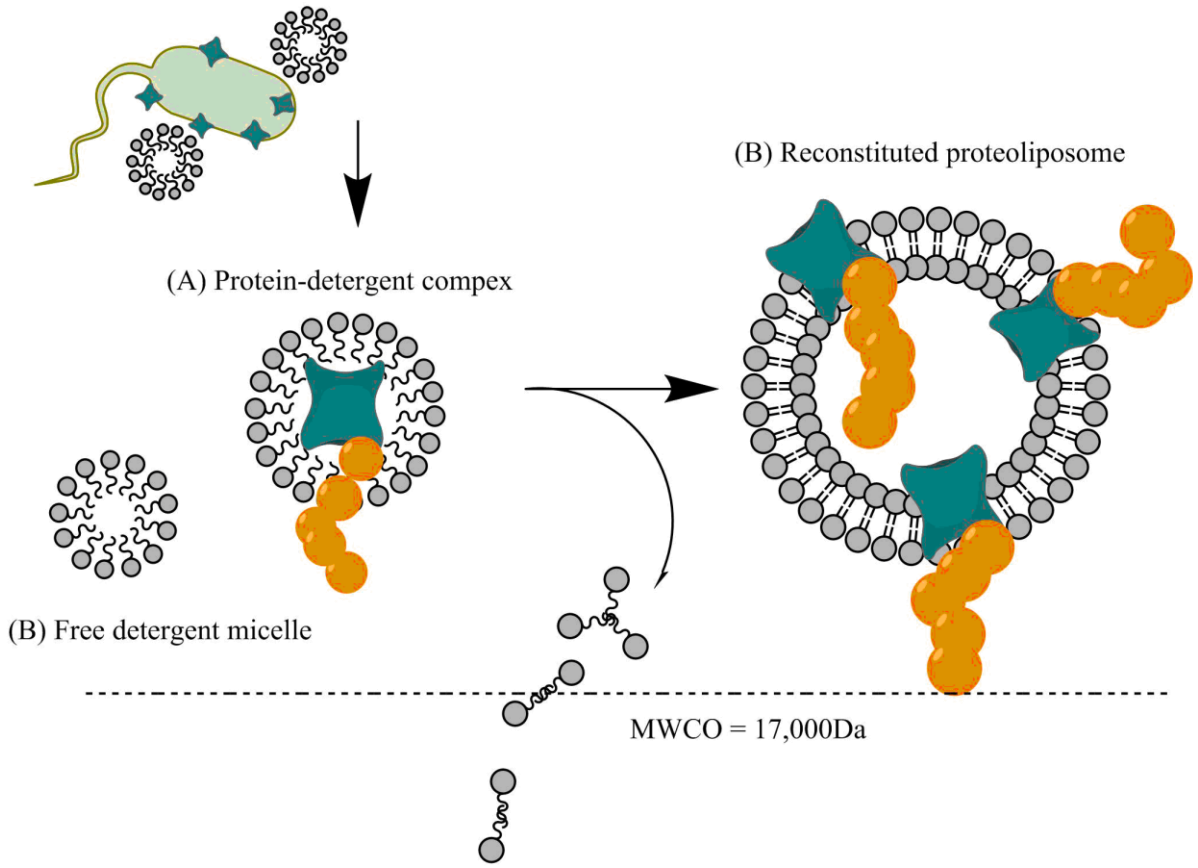


Figure 3.3 Membrane mimetics used for membrane protein solubilisation and reconstitution. Detergent are used to extract recombinant BamA from the *E. coli* membrane and protect the hydrophobic β-barrel from aggregation in detergent micelles (A). Dialysis is used to remove the detergent in free micelles (B) in the presence of synthetic liposomes to reconstitute the protein in native-like bilayer (C).

3.2.5 Application of detergents in membrane protein preparation

According to Equation 3.3, the work done to maintain the interface between two phases is proportional to the change in the surface area of the boundary. In both native membrane extraction and targeted recombinant expression systems, a buffered detergent solution is applied to isolated cell membranes to disrupt the lipid bilayer and release the membrane proteins into

solution (Figure 3.3). Lipids and proteins are then solubilised in water to form mixed lipid-detergent and detergent protein-micelle complexes.

In recombinant over expression systems, cytosolic protein aggregates are formed as inclusion bodies which can also be more easily disrupted using detergents and additional mechanical force from various techniques i.e. aspiration, vortex, or sonication. Depending on the number of exposed residues, larger membrane proteins typically present a larger hydrophobic surface thereby increasing their interfacial surface area. For this reason, a high concentration of detergent may be required, as the hydrophobic surface area increases, and they become more prone to aggregation. Furthermore, the amount of detergent used should be adjusted based on the protein concentration as spontaneous aggregation will occur more rapidly and be more difficult to reverse in certain conditions. Detergent buffer solutions used for solubilisation should also be composed of preformed micelles, typically for a solution made up at least three times stronger than the quoted CMC [90].

Initially, additional time and work is required for buffer preparation as both dehydrated solid powders, and detergents solutions in the liquid phase must be dispersed throughout the bulk solvent phase. This is best achieved through gentle mixing and moderate use of heat. However, vigorous shaking can quickly lead to excessive foaming and bubbles which become trapped on the top surface between the liquid phase and air. A concentrated detergent phase is then formed on the surface of the solution. This aspect of preparation may lead to poor yields of solubilised protein due to a heterogeneous mix being formed and should be noted before preparing detergent solutions. Furthermore, at higher concentrations, detergents have chemical properties which will disrupt protein folding leading to a, potentially irreversible, loss of tertiary structure. For this reason, successful membrane protein extraction protocols are usually developed after a series of optimisation steps which find the right balance for maintaining the highest concentration of the target protein in a soluble monomeric protein-detergent complex without causing denaturation or wasting materials.

3.2.6 Non-detergents membrane mimetics

The factors explained above are important to consider when designing protocols for handling membrane proteins as reducing the work required with the correct detergent, at an appropriate concentration, can ease the whole process. However, although detergents are the most commonly

applied mimetics in structural studies of membrane proteins, other non-detergent surfactants and novel molecules are also used. For example, styrene maleic acid (SMA), which has previously been used to extract α-helical inner membrane proteins, forms a co-polymer with the native lipids while retaining the native structure by removing entire patches of membrane. This approach yields complexes referred to as SMA lipid particles or SMALPs, which includes the membrane protein surrounded by a ring of native lipids. SMALPs have been used to investigate α-helical membrane protein structure with electron microscopy after extraction and purification directly from the membrane bilayer [91]. However, they have not yet been shown to be effective for extraction of β-barrels in outer membrane extractions, perhaps due to the charged polysaccharide molecules present in the native outer membrane environment. Furthermore, the effect of styrene maleic acid in further crystallisation stages has not been yet been found in literature, possibly as SMA-lipid particles could restrict protein-crystal formation.

3.2.7 Dynamic Light Scattering

Once a membrane protein has been isolated, the solution should be analysed for quality assurance purposes. Dynamic light scattering is a technique capable of determining the size distribution of small particles like protein-micelles in a liquid suspension by measuring the translational diffusion coefficient of molecules undergoing Brownian motion (Figure 3.4). Monochromatic light is scattered in all directions by small particles, so long as the particles are relatively small compared to the wavelength of the light. Although coherent and monochromatic light from a laser is used, scattering intensity of the emergent light fluctuates with respect to time. This is caused by heat and the particles in solution undergoing Brownian motion [92]. The scattered light will therefore be subject to interference caused by surrounding particles which move at different speeds, depending on their size. The intensity of these fluctuations contains information which can be used to calculate the time scale of motion by the scattering particles and processed with an autocorrelation function to determine a size distribution profile.

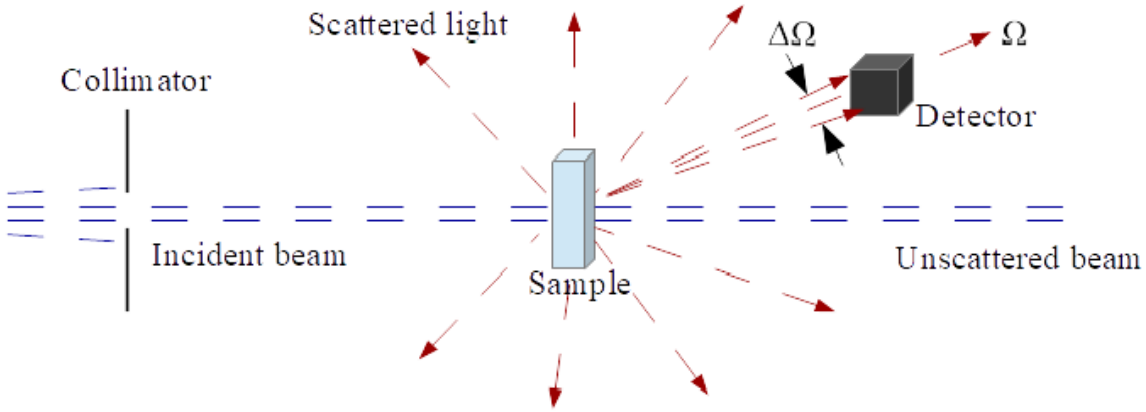


Figure 3.4. Schematic diagram of dynamic light scattering apparatus. An incident beam of light is scattered in all directions by the sample cuvette at a fixed intensity (Ω). The liquid inside the cuvette contains molecules which undergo size-dependant diffusion according to Brownian motion, which causes some of the light to fluctuate - the intensity of which is measuring (ΔΩ) at a photodiode detector and used to determine the size of molecules in the solution.

When a membrane protein is solubilised by a detergent and purified, it is desirable for a uniform homogenous protein-detergent-micelle complex (PDC) to be detectable as the dominant size, close to the size of the micelle. More importantly a PDC will appear with a dominant single-peak radial distribution, indicating the protein is folded. A membrane protein which is insoluble or misfolded in the chosen detergent buffer will appear with a complex radial distribution and at a much larger radii as a result of amorphous aggregation [93].

3.2.8 Atomic Force Microscopy

Atomic force microscopy (AFM) is an additional technique which is useful for characterising membrane proteins after reconstitution. The atomic force microscope is a type of scanning probe microscopy instrument which is used to measure particle size and size distribution analysis, perform single-molecule force measurements, and generate high-resolution surface topographies. AFM was developed by G. Binnig and H. Rohrer, who combined the principles of scanning tunnelling microscope and the stylus profilometer [94].

AFM works on the principle of measuring contact forces or *touch* rather than light, as in a traditional optical microscope, by using an ultra-sharp microfabricated tip which is brought into close contact with the surface to be imaged. As the tip approaches the surface, the electro-static force between the two objects causes it to deflect towards, or away from, the surface based on the properties of each material. The tip is attached to a cantilever which acts as a spring to sense the level of deflection, and a laser, which is reflected off the tip and directed onto a set of photodiodes, is used to measure displacement. The laser is aligned to hit the centre of the photodiode set so that the magnitude of deflection can be measured from the intensity of light reaching each individual diode. The height of the sample is adjusted by modulating a piezoelectric crystal which the sample rests on, in response to the signal from the photodiode set.

A PID (proportional-integral-differential) controller is used to modulate the sample height and maintain a relatively constant force between the tip and sample, according to user defined, or automatic, gain setting. This is known as contact mode. Alternatively, non-contact mode, also called peak-force tapping mode AFM, forces the tip to vibrate with high-frequency oscillations, tuned to the natural resonance of the cantilever according to the properties of the spring-like tip. This set-up maintains a peak intermolecular distance between the tip and the sample without excessive force which could otherwise cause damage to more delicate biological samples, and with less risk of damaging the tip. The topology of the sample is generated from a 3D reconstruction

of the digital signal through the feedback system. With the correct set-up, AFM can obtain images at extremely high sub-Angstrom resolution, over a thousand times greater than the diffraction limit of traditional optical light devices [95].

Electron crystallography and AFM allow the study of membrane protein structure, surface properties and even conformational dynamics, at sub-nanometre resolution using two-dimension crystals [96] . 2D crystallisation is beneficial in such studies and for achieving higher resolution with AFM for several reasons. Firstly, multiple instances of the same protein structured in repeats allow for greater sample rates and improved reliability. Conformational changes can also be observed occurring in large populations of hundreds or even thousands of units, rather than in a single molecule. Secondly, a unit in a planar crystal lattice is less prone to thermal drift and stylus effects which can cause sample deformation and poor resolution. Oscillating non-contact tapping mode AFM may be used instead to image single biomolecules with less force, as the tip only reaches a maximum force at the height of the tapping motion, unlike constant force mode. However, this method is difficult to implement practically in liquid. High resolution images can still be generated using this mode of AFM in suitable conditions and with the appropriate gain settings but the availability of 2D crystal arrays in scanning probe microscopy remains desirable.

Atomic force microscopy has been used in other areas of research to investigate surface properties such as elasticity and stiffness and acquire high resolution images of surface topography, but it is particularly suited for in-situ studies of reconstituted membrane protein systems. Imaging can be performed at normal environmental conditions, using physiologically relevant buffers and without the need for freezing or dehydrating samples, allowing dynamic conformational changes involved protein-ligand interactions to be investigated more thoroughly. Moreover, AFM addresses many of the difficulties in studying membrane proteins by facilitating investigation in a lipid bilayer environment, retaining a native-like state and AFM has already been used to study a wide range of membrane protein structures and their interaction with various ligands [97]. Examples include induced conformational changes of reconstituted membrane proteins studied in systems such as bacteriorhodopsin with light [98], OmpF and OmpG with H^+ protons [99] and NhaA with Na^+ (sodium) ions [100]. The latter three are microbial outer membrane ion channels which control the ionic strength in the periplasm, which have been reconstituted to form 2D crystals in a synthetic lipid bilayer. Remarkably, conformational changes in the surface could be monitored in real-time by AFM by adding a strong ionic solution over the sample which contained the respective ligand appropriate to induce activity in each system.

3.3 Results

Initially a high level of protein aggregation was detected in the first expression system used (D. Griffin/N. Oldfield, 2012). Full length *Cj*BamA had been cloned with a C-terminal hexa-histidine affinity tag and over-expressed in *E. coli*. High expression levels were achieved by inducing expression using 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) in 500 mL batch cultures grown in LB media at 37°C to encourage the formation of cytosolic protein aggregates as inclusion bodies. Inclusion bodies were isolated after cell lysis using a repeated washing and centrifugation cycle, denatured in 8M urea and purified by anion-exchange chromatography on an AKTA-FPLC. Relatively pure BamA (above 95%) was confirmed using SDS-PAGE at approximately 83kDa. Refolding was then attempted in 0.5% (w/v) *N*-dodecyl-*N,N*-dimethyl-1-ammonio-3-propanesulphonate (SB-12), 20mM Tris–HCl (pH = 8.0) and 1M NaCl by 10-fold rapid dilution into a 10 mL reservoir and left over 72 hours at 37°C as described previously [77]. Secondary affinity chromatography was then used to target refolded protein with immobilised metal affinity column chromatography, however this proved unsuccessful, as most of protein was found in the flow-through fraction after loading. To confirm the presence of the polyhis tag, Western blotting was used after SDS-PAGE and confirmed the availability of the tag at the appropriate size.

3.3.1 Particle size analysis by DLS and AFM

Production of pure BamA by over expression, denaturation and ion-exchange chromatography was confirmed by SDS-PAGE, as described above. However, attempts to subsequently further purify the protein after refolding using immobilised metal affinity chromatography were unsuccessful in preliminary trials. The formation of protein aggregates during refolding was investigated using dynamic light scattering and atomic force microscopy. After refolding, the protein was then concentrated using a 30 kDa molecular weight cut-off ultrafiltration device (Millipore) and reconstituted in 1,2-dimyristoyl-s*n*-glycero-3-phosphocholine (DMPC, Avanti Polar Lipids) by slow dialysis at a 100:1 molar lipid to protein ratio.

Particle size analysis was performed using dynamic light scattering (Malvern Instruments) after samples were filtered using a sterile 0.45 μm filter device to remove large particulate matter. Analysis of a control liposome solution showed that a dominant peak particle-size was present with an approximate molecular weight of 175 kDa (Table 3.1 and Figure 3.5). However, significantly larger particles were found in the reconstituted protein samples that were not present

in the liposome control solution. In a control liposome preparation, the dominant fraction is observed at an approximate weight and size of 663 kDa and 175 nm, respectively. In the solution with BamA, up to half of the particles have a much higher molecular weight exceeding 200,000 MDa with an approximate size of 314 nm.

According to literature, unilamellar DMPC liposomes can vary in sizes from 5 – 50 nm [101]. This suggesting that the slow-dialysis method used has formed large multilamellar vesicles (MLVs). These findings therefore do not exclude some level of protein reconstitution however, the formation of MLVs would the determination of successful protein reconstitution in this manner.

Table 3.1 Particle size of analysis of proteoliposomes

	Liposome control			BamA Proteoliposomes		
Peak	*Area (%)*	*Rh (nm)*	*MW (kDa)*	*Area (%)*	*Rh (nm)*	*MW (kDa)*
1	3.2	0.23	0.11	0.7	0.78	1.82
2	0.7	5.70	202.48	3.0	7.25	357.47
3	**84.1**	**175.38**	**6.63×10^5**	17.3	69.82	7.52×10^4
4	12.0	1624.34	127×10^8	**47.1**	**314.60**	**2.64×10^6**

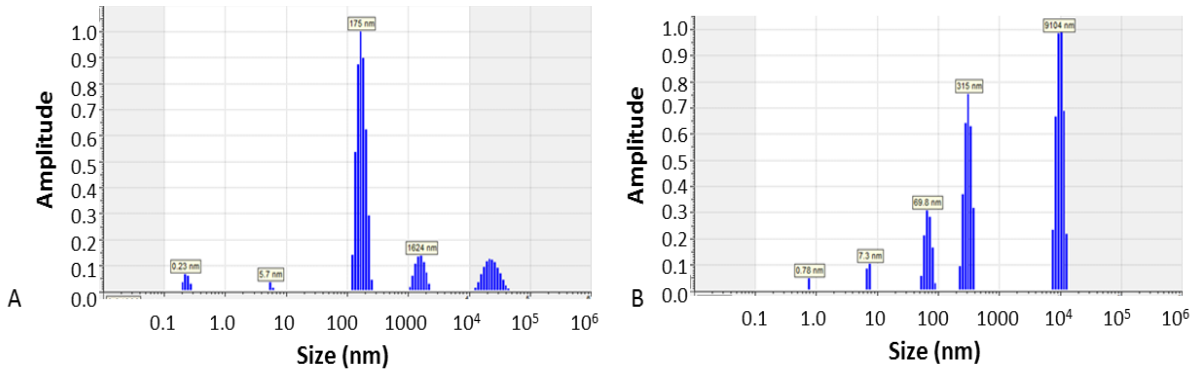


Figure 3.5. Particle size analysis by dynamic light scattering of reconstituted BamA proteoliposomes. Reconstitution of refolded BamA in DMPC liposomes was performed using slow-dialysis to remove detergent. In the liposome control sample (A), a dominant peak of approximate 175 nm is observed by DLS (Table 3.1). With BamA (B), DLS indicates that larger particles have been introduced suggesting a high polydispersity arising from suspended protein aggregates with sizes up to 2.64 MDa.

To further investigate if any protein reconstitution had occurred, atomic force microscopy was used to confirm the presence of large aggregate material, after reconstitution and flattening on a freshly cleaved mica surface. The proteo-liposomes were dispensed on the surface and allowed to flatten for up to 4 hours. A small amount of buffer was added at regular intervals to prevent the sample from drying and kept in a sealed petri dish. Peak-force tapping mode was then used to collect images on a Bruker-Nano MultiModeV atomic force microscope (Figure 3.6).

The images collected showed that flattened bilayers were formed at an approximate height of 5 nm, close to the predicted theoretical membrane height. Some areas appeared to feature stacked membranes, further indicating that MLV formation had occurred during liposome preparation. Again, the success of protein reconstitution was difficult to determine by AFM. The presence of large aggregates on the lipid bilayer surface, up to 100 nm across, were observed however smaller particles were also seen distributed across the surface which may have been reconstituted BamA. Further attempts to image smaller areas of these sections led to sample degradation and individual structures were therefore not able to be resolved in this manner.

3.3.2 Cloning, low temperature expression and purification of BamA

Redesign and optimisation of the expression and purification method was done to address the problems associated with protein aggregation shown previously. Based on recent trends in literature which described low temperature cultures and limited IPTG concentration induction for membrane protein production, an adapted system for expressing *C. jejuni* BamA was designed and used to successfully extract folded protein from the host *E. coli* outer membrane preparation, solubilised in detergent micelles. Firstly, the *Cj*BamA gene was sub-cloned in a modified pET20b(+). The gene was amplified by PCR using Taq polymerase, followed by digestion using NcoI and BamHI restriction enzymes. The expression plasmid was prepared through transformation in competent *E. coli* BL21 by heat shock, followed by overnight culture, lysis and purification using a GenElute MiniPrep kit (Sigma-Aldrich, UK). The purified plasmid was then identically treated with restriction enzymes and ligated with the amplified gene using DNA ligase. After ligation and transformation by heat-shock into a suitable *E. coli* BL21 expression strain, the cloned plasmid was isolated and purified. Polyacrylamide gel electrophoresis were used to detect and confirm the presence of the ligated PCR product and pET vector.

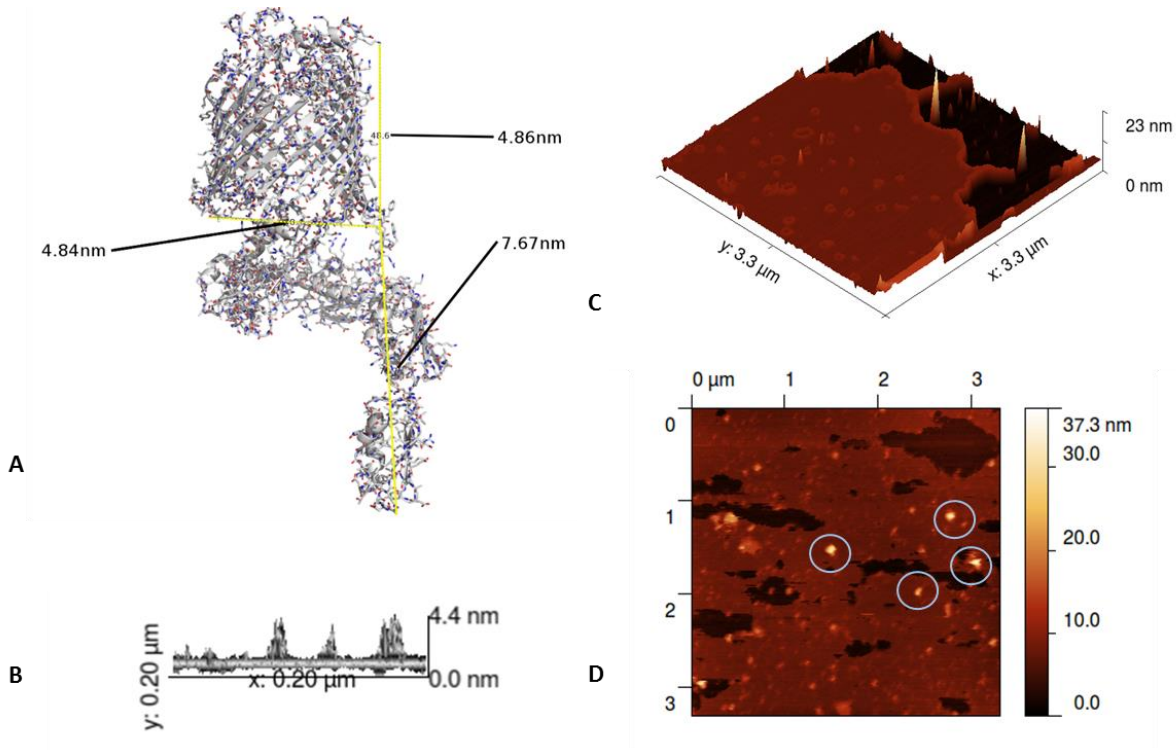


Figure 3.6 AFM images indicate the deposition of stacked DMPC bilayers of BamA-DMPC liposomes. Approximate dimensions of BamA monomer (PDB: 4K3B), measured in PyMOL, are: β-barrel domain; 5x5 nm, POTRA domain length; 7.67 nm (A). The larger particles seen with AFM therefore represent oligomers ranging from a hundred to a thousand monomeric units - indications of the formation of massive multimeric aggregates (B, C, D).

Whole lysates were blotted using anti penta-histidine primary antibody for untransformed *E. coli* BL21, and two expression levels tested using 0.4mM IPTG at 20°C and 1mM at 37°C. Lower induction produces expression levels closer to native levels seen in the *E. coli* whole lysate (Figure 3.7). Elution fractions from IMAC, performed in TBS with 2% DDM, show a prominent band at the expected weight, slightly above the 75 kDa marker. Another less intense band is also observed below the 75 kDa marker.

The larger apparent weight band, which represent the fully unfolded β-barrel is thought to be caused by a band shift, whereas the lower band has retained the condensed β-barrel fold. Ideally, SDS and heat denaturation are used to completely unfold the protein for accurate size determination by SDS-PAGE, however initial attempts to boil samples caused the protein to precipitate; an affect which will be discussed further in later sections (please see Chapter 5).

3.3.3 Detergent buffer screening

After purification of *Cj*BamA with the pET20b expression system was confirmed as described above (Section 3.3.3), an ultracentrifuge dispersity sedimentation (UDS) assay was used to assess the suitability of various detergent solutions for preventing aggregation during extended storage periods [102]. Furthermore, these results would be used to determine the use of appropriate detergents for crystallisation trials. Elution fractions were pooled after primary IMAC and buffer exchanged to the detergent solutions using 100 kDa molecular weight cut-off ultrafiltration spin concentrators. Although the approximate theoretical size of the *Cj*BamA-10xHis was calculated as 85kDa (ProtParam, ExPASy) a larger membrane cut-off size was used to effectively remove additional free detergent micelles while retaining the larger protein-detergent micelle complexes.

The buffer exchange step was repeated three further times with fresh buffer to remove at least 99% of the initial 0.1% DDM solution used during purification and left overnight at 4°C to allow aggregation to occur. A small amount was taken and mixed with SDS loading buffer before the solutions were spun at 260,000 g to remove aggregated material and particulate matter. The results (Figure 3.7, Box C) show that Tris buffered saline solution (200mM NaCl, 50mM Tris, pH 7.5) with 0.1% (w/v) SB12 and 0.8% (w/v) CHAPS maintained protein solubility in solution.

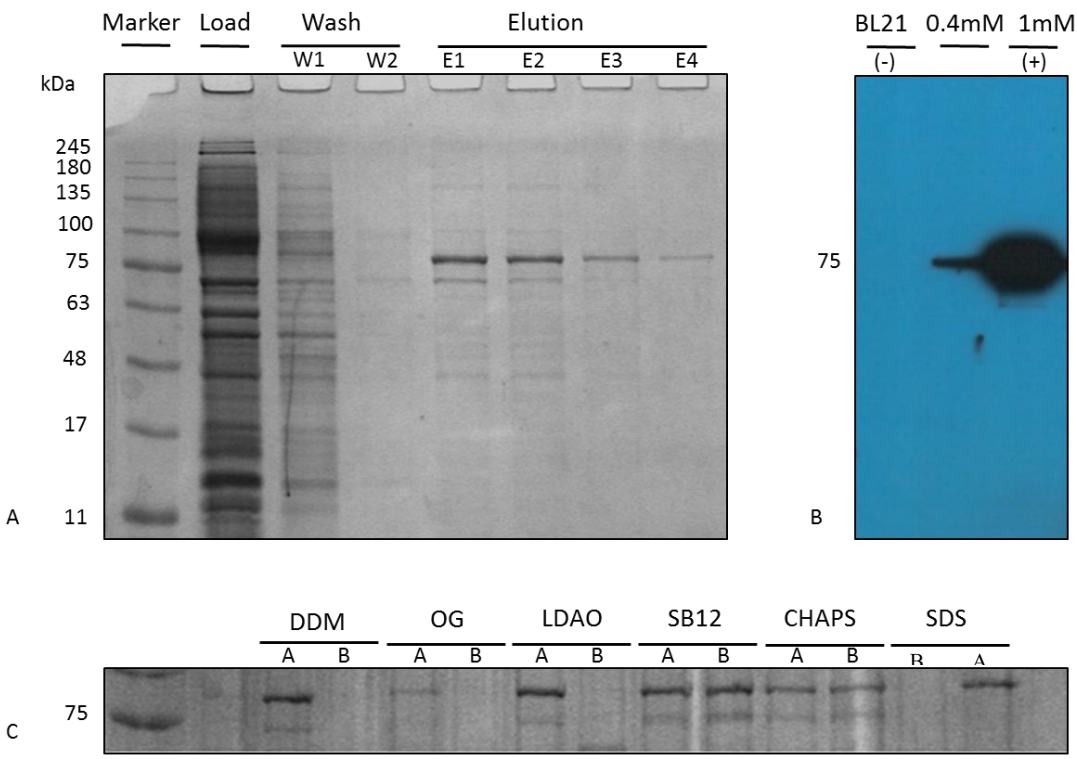


Figure 3.7 Purification of *Cj*BamA; detection using SDS-PAGE, Western blotting and detergent screening with ultracentrifugation dispersity assay. SDS-PAGE confirms purification of BamA after primary immobilised metal affinity chromatography (IMAC) in lanes E1-E4 (frame A), while Western-blotting confirms the presence of a poly-histidine tag and IPTG induction (frame B). An ultracentrifugation dispersity sedimentation assay was used to determine suitable detergent buffers for solubilising BamA (frame C, A: before, B: after). Only SB12 and CHAPs solutions maintained solubility after 16 hrs at 4°C.

In all other detergent conditions, the protein was recovered as sediment from solution by ultracentrifugation, indicating that the protein had aggregated, and the detergent solution used was unsuitable for maintaining membrane protein solubility in this case.

In each lane for sample before ultracentrifugation (Figure 3.7, C), two bands can be seen which represent the folded and unfolded state of the β-barrel - the lower apparent molecular weight species represents the folded state, while the higher band is the unfolded protein. This is observed in all detergents, except SDS. This observation suggests that the denaturing SDS used in the PAGE sample buffer is only sufficient to denature a fraction of the BamA molecules and explains why the entire fraction of the protein appears completely unfolded at the correct apparent molecule weight in the 0.5% SDS solution.

From the gel image, certain concentrations also appear to be lower, despite concentrations being normalised after buffer exchange, based on 280 nm UV-spectroscopy readings. This effect could be explained by competitive binding of the Coomassie G-250 dye used in band visualisation. However, as an excess of Coomassie was used, this would have more likely been due to the formation of larger, mega-Dalton scale molecular weight protein aggregates which did not enter the resolving gel – an affect also seen in other later studies (please see Chapter 5). Due to the use of binary detergent solution in the final stage i.e. the screened detergent with SDS, it is difficult to fully determine the quantitative levels at which the primary detergent of choice can solubilise *Cj*BamA. Nonetheless, the results clearly show that SB-12 and CHAPS promote solubility as the protein is still found in the supernatant after high speed ultracentrifugation.

3.3.4 Crystallisation and X-ray diffraction crystallography

Followed confirmation by SDS-PAGE, protein samples were then prepared in 0.03% DDM as described previously, from 2 litres of LB media and purified by primary IMAC and Size Exclusion Chromatography. Preliminary crystal trials were previously set-up at Diamond Light Source (I. Morales, 2013), Imperial Membrane Protein Lab using the MemGold2 kit (Molecular Dimensions, UK) and a Mosquito pipetting robot, in sitting drops using protein prepared from overexpression. No crystals were formed after 46 weeks. Further trials were then performed using the MemPlus screen, with 48 additional conditions targeted for β-barrel proteins. For these trials, the protein solution was prepared from 10 litres of Terrific-Broth, extracted from the outer membrane fractions, and purified by IMAC and SEC-GPC (Figure 3.8). Two detergent buffers were then

prepared, and the protein was buffer exchanged to 0.8% CHAPS and mixed with DMPC-CHAPSO bicelles.

Prior to mixing with bicelles, protein samples were concentrated to at least 8.0 mg/mL. Bicelles were prepared using a 2.7:1 ratio of DMPC to CHAPSO by repeated cycles of heating and cooling on ice followed by vigorous mixing using a vortex. The concentrated protein solutions were then mixed with the bicelle solution and kept on ice until aspiration. 40 nL of BamA protein-bicelle mixture was carefully spotted onto a hanging-drop cover sheet and an equal amount of precipitant layered on top. The sheet was inverted and sealed in a 96-well, SBS format flat bottom plate with 48 unique precipitant conditions (MemPlus screen, Molecular Dimensions, UK).

Small needle-shaped microcrystals were formed after 48 hours in several conditions. It was noticed that in most of the conditions with crystals, the precipitant solution included a high concentration of sodium chloride. However, they did not appear to have the cubic structure of a NaCl crystal, instead displaying long mono-clinic, rhombic like shapes. To confirm their composition, the largest crystals were fished and shot at the micro focus beamline, B21 (DLS facility, Oxfordshire). However, of the crystals studied, all diffraction patterns showed uniform concentric ring patterns characteristic of extended alkane chains (Figure 3.9).

3.4 Discussion

High purity samples of *Cj*BamA were initially obtained at yields estimated above 100 mg per litre of LB media. To achieve this, the protein was extracted from inclusion bodies formed by IPTG induced over expression, and therefore required subsequent denaturation and refolding. It was observed though, based on preliminary results from dynamic light scattering and atomic force microscopy experiments, that this system was unsuitable for producing a homogenous solution of monomeric protein-micelle complexes. Furthermore, the protein was initially purified via ion-exchange chromatography however subsequent polishing with metal-affinity after refolding was ineffective.

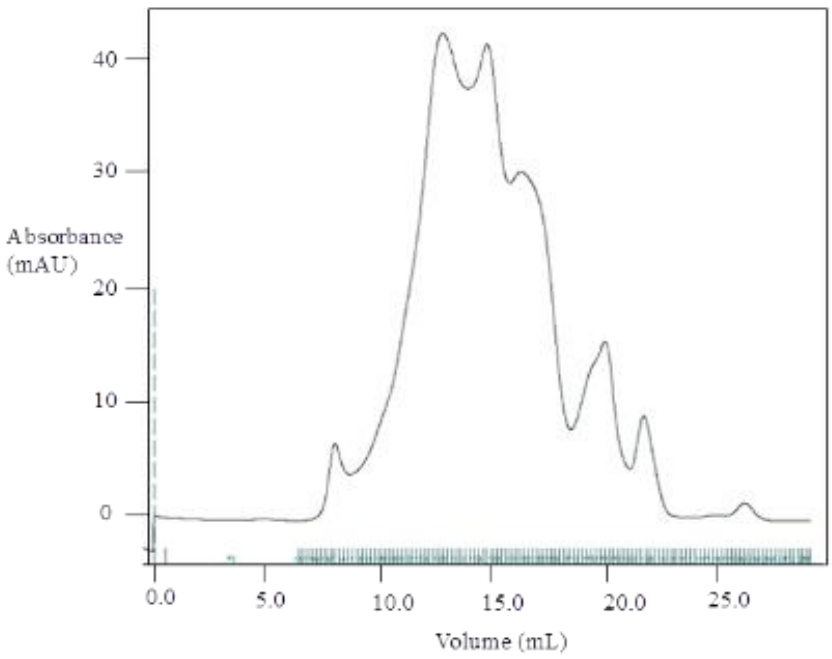


Figure 3.8. SEC-GPC trace shows multiple size *Cj*BamA species after primary IMAC. Size exclusion gel permeation chromatography was performed using a Superdex-200 resin column on an AKTA-Prime Plus FPLC system. A dominant fraction was eluted after 12 mL, although a number of other peaks are observed in the trace. This result shows that after primary IMAC the solution contains a size-heterogeneous mixture of proteins which could indicate misfolding or oligomerisation.

3.4.1 Affinity tag placement disrupts BamA folding

Previous reports of successful BamA purification which led to the acquisition of high-resolution structural data, as well as preliminary template-based homology modelling results (Chapter 3) suggested that the C-terminal region was likely to be masked in a detergent micelle and hence an unsuitable site for an affinity tag. For this reason, the first aim in redesigning the expression system was to create a new plasmid with an N-terminal histidine tag. This was achieved by using the pET20b plasmid which also features a pelB signal peptide recognised by the *E. coli* host SecYEG pathway and a TEV cleavage site located after a 10X poly-histidine tag for subsequently removing the tag after purification. The plasmid was then successfully transformed into a suitable BL21 plysS expression strain.

The original construct vector used in the over-expression system was designed with a C-terminal hexa-histidine tag (*Cj*BamA-His6). It was soon realised though from structural reports of BamA and other Omp85 proteins, that the C-terminal region was unsuitable for such modification. The addition of six, charged residues at this site would likely restrict proper folding of the β-barrel, and if correct folding had occurred the tag could also be masked or buried by the detergent. For this reason, the expression system was redesigned to create a new plasmid with an N-terminal decahistidine tag and TEV cleavage site which would be accessible for affinity chromatography (His10-TEV-*Cj*BamA). Based on recent trends in the literature which describe low temperature cell culturing and limiting the concentration of IPTG for BamA production, an adapted system for the expression of *C. jejuni* BamA was implemented to extract mature protein from a host *E. coli* outer membrane preparation using this construct.

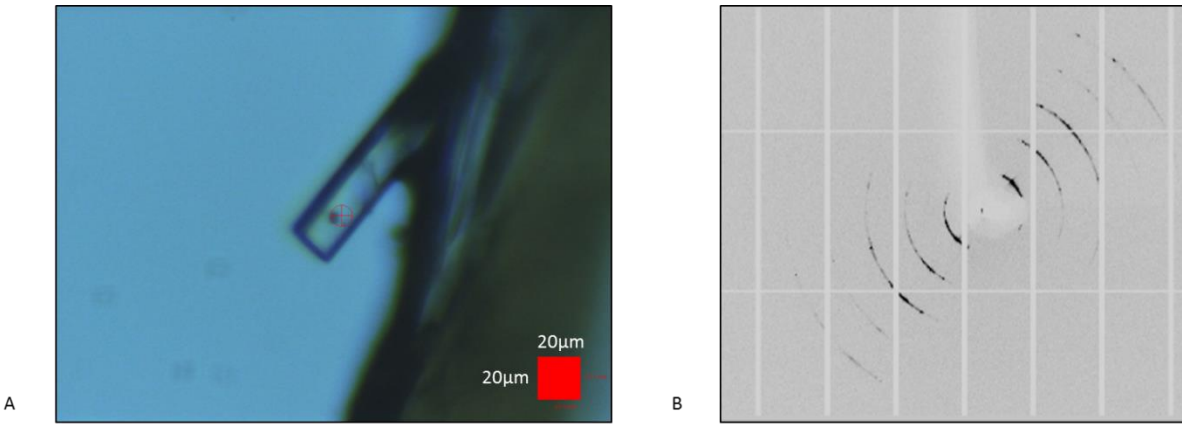


Figure 3.9. False-positive crystals and the X-ray diffraction pattern from a *Cj*BamA DMPC-CHAPSO bicelle crystallization trials. Lipids crystals were formed with the MemPlus crystallisation screen and purified BamA in C_8E_n. A number of crystals were fished (A) however concentric ring patterns – a hallmark lipid and detergent crystals – were observed (B).

3.4.2 Localisation of *C. jejuni* BamA limits growth in *E. coli*

To target the recombinant *Cj*BamA protein to the host-cell outer membrane, a pelB leader sequence was needed. The modified pET20b plasmid was therefore used to incorporate a signal peptide recognised by the *E. coli* host Sec system for secretion. The plasmid was then transformed into a suitable *E. coli* BL21-plysS expression strain, and the resulting organism called pMH1.

It was initially observed that growth on LB-agar plates and in liquid media at 37°C was poor for the pMH1 strain, with optical densities measured at 600nm reaching no greater than 0.5 after 16 hours of growth. Furthermore, preliminary cell culture tests indicated that 1 mM IPTG, typically used in over expression systems, was also a limiting factor for cell growth. It is thought that temperature and IPTG therefore lowered the growth rate due to a toxic effect caused by periplasmic aggregation. Non-native periplasmic protein secretion has previously been recognised as a source of toxicity and poor growth [103], as it can disrupt the membrane environment and adversely affect cell morphology. Alternatively, when at higher concentrations, then BamA from Campylobacter may have upset the operation of the *E. coli* BAM system and interfered with periplasmic OMP trafficking, further affecting cell growth rates. Furthermore, over expressions of a pelB target could have inundated the cytoplasmic SecB chaperones and prevented proper assembly of the outer membrane.

Temperature and IPTG concentration were therefore optimised between 16 – 27°C, and longer growths were carried out at a reduced temperature of 23°C with up to 0.4 mM IPTG. This balance of expression and growth allowed high cell-density liquid cultures to be grown with good OD_{600} measurements above 2.0, achieved after 16 hours growth. Further purification by metal-affinity chromatography was also achieved after several weeks of trials using a variety of resins including Nickel-NTA, Nickel-IDA and different column sizes. The methods developed were then used to prepare *Cj*BamA for crystallisation trails and applied in further studies shown later (please see Chapter 5).

3.4.3 Yield optimisation for crystallisation

Material costs, time and energy were reduced by omitting FPLC in favour of manual batch purification using loose Ni-NTA IMAC resin. This saved significant amounts of detergent-buffer as the resin volume needed was much less compared that need for a column method, and reduced preparation time. Furthermore, the use of lower resin volumes is attributed to higher purity and

better yields from the process. This was because the product expression was the limiting factor in this system, which means that excessive chromatography resin had a negative impact on yields. Although counter-intuitive, using less resin generated higher yields. This is explained as the binding capacity of resins is usually not met by the level of target protein available from membrane extractions and is normally done to ensure no target protein is lost. However, this is only useful in an overexpression system when the target protein is available in much higher levels compared to the levels of non-specific impurities. As such, if the total binding capacity is higher than the amount of target protein available, then a higher level of impurities will be found in the elution fractions. The novel tip-chromatography method takes advantage of this theory, essentially using the same principle of the manual batch method employed here but also being viable for automation with liquid handling robots [104].

Despite the difficulties encountered here, protein refolding methods have since been used to solve the structure of BamA obtained by over expression [105]. Once the protein aggregates were isolated, a chaotropic solution such as 8M urea was applied to disrupt hydrogen bonding networks and therefore effectively denature the BamA. After aggregates have been solubilised, the denatured proteins are trapped in a high energy, unfolded state which are restricted from aggregation by the addition of chaotropes or detergents. Protein refolding can then be performed by rapidly dispersing the denatured molecules into another detergent buffer. In principal, the protein rapidly refolds as it falls to a lower energy, more stable state, while detergent molecules spontaneously mask the exposed hydrophobic surface, trapping the protein in a micelle and suppressing aggregation. However, despite being shown to be practically feasible due the location of the poly-histidine tag, in our over expression system, these refolding methods did not work. The success of refolding is measured as the fraction of protein which forms into monomeric protein-detergent complexes with the native fold, over the amount of total protein. A refolding protocol typically takes several days though, and success is dependent on many factors including the choice of detergent, detergent concentration, buffer properties, reaction time and temperature, amongst other. Here we have shown that high yields of protein could be obtained through these methods, but the procedures took significantly longer and were more technically challenging to execute and monitor, compared with the methods for targeted membrane extraction.

3.4.4 Application of high-end structural techniques

Although numerous BamA structures have already been published, expanding the database of membrane protein structures is central to understanding how the membrane functions alongside

the intracellular components of` the cell. If essential components are missing, then it is not possible to form an accurate structural or mechanistic model. Furthermore, each crystal provides new insights into different conformations that the protein can adopt.

While several full-length BamA crystal structures have become available, due to the relatively low sequence identity of the *C. jejuni* BamA gene identified through homology, we thought determining its structure experimentally would further our understanding of structural relationships within this essential family of proteins. X-ray crystallography diffraction is one of the most commonly used methods to build structural models of biological macromolecules and yet it is still widely regarded as a challenging undertaking, with many inherent difficulties encountered here.

Although the initial stages of cloning, expression and purification follow those used for soluble proteins, the non-trivial inclusion of detergents used to maintain solubility for normally insoluble molecules presents significant limitations [106]. Nonetheless, significant progress has been made in the field in recent years, namely due to the development of better automation, and novel tools, including screening by rational design, a wider range of commercially available detergents, lipidic cubic phases and bicelle crystallisation systems, making membrane protein crystallisation much more feasible [107].

The development of other scale down screening tools such as the UDS assay, used here to identify suitable detergents for handling *Cj*BamA, are also leading to improved membrane protein preparation protocols. The application of more sophisticated genetic engineering methods, which are also cheaper and easier to use have been shown to be successful in this case for enhanced expression, isolation, and purification of *C. jejuni* BamA. Since this work was conducted, additional BamA structures have been solved by XRCD from solutions prepared using both over-expression and membrane extraction methods. After trialling both, each were found to have their own respective merits and the most appropriate method would be dependent on the material costs and equipment available.

In any case, membrane protein preparation is an expensive undertaking with no guarantees of success. Various methods are available but due to the high material and energy costs associated with structural studies, it is important to optimise preparations before scaling-up. By choosing the appropriate methods, optimising preparations is an important step towards improving the quality

of samples and increasing yields. The cost benefits and reduction of waste are also important factors which could be further examined and assessed to provide positive indicators in future process design. Ultimately, better design circumvents the need to perform a tedious trial and error style process optimisation, saving time and reducing material costs.

Using the methods developments made in this section, efforts towards solving the structure of *C. jejuni* BamA can be taken forward in future work. In the next chapter, computation methods are used to predict the structure of *Cj*BamA using template-based protein homology modelling software. Later, more *in vitro* samples are further studied using circular dichroism to investigate the secondary structure profile and thermal stability of purified BamA protein samples to provide further insight that could aid future attempts at high resolution structural studies.

Chapter 4. Template based protein structure prediction of BamA and BamD proteins

4.1 Introduction

Initial attempts to obtain a high-resolution structure of *C. jejuni* BamA by X-ray crystallography proved difficult. Although significant progress was made to optimise sample preparation, due to the availability of homologous structures of *N. gonorrhoeae*, *H. ducreyi* and *E. coli* BamA proteins, comparative protein modelling was also considered a suitable additional method to predict the structure of the target *C. jejuni* protein, while further progress towards experimentally solving the structure was underway. In this section, computational methods which incorporate both existing experimental findings and structural bioinformatics tools are used to predict the structure of *C. jejuni* BamA. Furthermore, the discovery of a previously unlabelled gene, identified as the absent BamD component in *C. jejuni,* is reported in this organism and other *Campylobacters*. These models are used to build a homologous Bam system with two essential components that mirror the pathway seen in *E. coli* and provide a structural basis for further investigating the behaviour of these proteins using molecular dynamics simulations described later (please see Chapter 6).

4.2 Background and theory

Proteins are biological macromolecules built from constituent amino acid residue building blocks. Peptides - short chains of amino acids which are bonded into larger structures known as polypeptides or proteins - have many functional and structural roles in catalysing and modulating biochemical reaction pathways. Encoded for by DNA, the process by which proteins are synthesised within the cell has previously been described through the communication of information [108]. Since the discovery of DNA, genetic analysis has provided insight the origins and relationships between all organisms. These are traditionally identified through phenotypic traits and used to better understand biological function. However, the chemical mechanisms alone which lead to the information paradigm cannot fully describe these processes.

Functional relationships within biological systems are governed by catalytic processes arising from complex spatial rearrangements occurring on the protein level, many of which are still difficult to observe using traditional experimental methods. Nonetheless, deciphering the intermediate stages of the central dogma of molecular biology is only possible by incorporating various research

methods from a range of disciplines. Specifically, the biophysical theory of how protein structure and function evolved alongside organisms is still underdeveloped. Due to its ancient origins though, BamA and the β-barrel assembly machinery present an ideal case to investigate these areas in numerous organisms found in different domains of life.

4.2.1 Evolution of Protein Structure and Complex Assembly Machinery

Evolution is traditionally viewed at the organism level. Indeed, variation in characteristic traits such as behaviour and appearance can more easily be explained by adaptation to environmental conditions, and before development of the modern synthesis, which established a framework between evolving genetic elements and traditional views on natural selection, there was no other explanation of how molecular components could lead to such variation in nature [109].

Nonetheless, modern synthesis tries to explain how molecular biology works through progressive mutations in genetic elements which can lead to beneficial traits in the organism [110]. The stronger or *fitter* organism is better adapted and therefore should have a better chance of survival. In turn, it is more likely to carry its genes on and, more importantly, pass them on to their progeny. Over many millions of years this process can lead to the massive variation we see in biology today. But how is it that such massive variations in genetics are seen while all organisms essentially perform the same basic functions?

The modern synthesis does not fully explain many of the phenomena observed in nature and additional theories regarding the early evolution of macromolecular structure are required. Central to these theories beyond the modern synthesis is the idea that the evolution of macromolecules occurs through complex networks of stochastic chemical reactions which have grown over billions of years [111] [108]. As the central dogma explains, the information in DNA is passed to proteins through a code which determines their structure. Recent theories for the evolution of macromolecules suggest though that to maintain function, structural properties must be constrained while the information can evolve freely. The majority of genes are therefore derived from replication events which are limited to constrain structure and enable conservation of extant enzymes. Some of these structures, which may be up to four billion years old yet have been shown to crystallise with remarkably similar structures despite variation in sequences [112] .

Nonetheless, although spatial constraints on structures must be imposed to preserve function, which is thought to be more conserved than sequence, some homologous protein sequences can

also fold in distinct ways [113] making it difficult to trace protein ancestry and predict the shape of new sequences.

Evidence of evolution for macromolecular assemblies such as BamA is also difficult to obtain with current methods. Microevolution, which occurs within a species, is known to take place in existing organisms, yet macroevolutionary changes are much more difficult to determine. As such, the origins of complex assemblies such as passive molecular motors like ATP-synthase, DNA-synthase and the BAM complex cannot yet be explained with much certainty [114] [115] [116].

4.2.2 Theory of structural macromolecular evolution

The evolutionary processes leading to the Bam and other assembly systems are generally explained by mutations that cause proteins to fold in new ways leading to gain of function. An example of this is a mutation in human blood clotting protein Factor VIII, a rare condition which happens in humans and which can be life-threatening [117]. Haemophilia, which leads to an excessively abnormal amount of bleeding can occur due to a deficiency in the blood clotting factors if they are not expressed sufficiently. Alternatively, an inactive structure of the Factor VIII protein can be expressed due to a specific genetic mutation. The protein can still fold but a single point mutation which affects the primary sequence causes the structure to be inactive. Both of these are considered evolutionary *disadvantages*.

Alternatively, a mutation which increases protein stability or perhaps heat tolerance, allowing an organism to survive in a wider range of conditions, would be considered an evolutionary *advantage*. Such changes are frequently observed and utilised in organisms such as *Thermophilus aquateqous* [118] which can survive at extremely high temperatures. There are two concepts of selection underpinning this selection pressure, both of which happen through changes in evolving macromolecular structures. Firstly, the changes which impart enhanced or new enzymatic functions by modifying the structure and secondly, changes which allow the structure to function in a wider range of conditions. Additionally, genetic elements have a vital role in rates of expression, so these also have their own implications on chances of survival.

Some mutations are synonymous which means they do not produce a change in the amino acid sequence which is coded for in the protein. For this reason, measuring homology using DNA alone can be misleading. Secondly, neutral amino acid variations can occur when a different residue

is coded for but causes no significant change on the structure and charge distribution of the protein. This happens as some amino acids share similar molecular size and charge properties and can therefore be substituted for one another without having any functional effect on the gene. Furthermore, amino acid substitutions occurring in or around an active site related to function could be deleterious whereas many substitutions occurring in other areas of the gene may cause inconsequential changes on the protein structure.

To overcome limited experimental methods and lack of sufficient information about early evolution events, computational solutions are used in place of empirical findings to provide data. To studies these events, probabilistic methods provide a framework and can be used to generate predictions not only about protein structure but also about function. Furthermore, protein evolution has already been simulated theoretically using fitness landscapes and population dynamics to identify dominant selection pressures on protein sequences and structures [119]. Nonetheless, understanding how each individual components within the cell arise to form a complete life-cycle still presents a massive problem [120].

4.2.3 Dynamic programming and template-based structure prediction

Dynamic programming algorithms are common place in modern computing, with applications in mathematics, economics and bioinformatics [121] [122]. They are used for structure and function prediction of proteins, and valuable because of the challenges for determining atomic resolution structures using experimental methods, as discussed previously. Essentially, dynamic programming is used as a method to detect homology between any two amino acid sequences and then available structures are used to build a model of the target sequence, without any need for traditional wet-lab experiments. A library of pre-determined structures is first required which ideally has a close homologue of the target protein structure, solved at high-resolution, although *de novo* methods exist to predict structure for unknown targets are also available [123].

Dynamic programming means using methods which compute a problem by breaking it into smaller, sub-problems - by storing the previous solutions - so when the same computation is needed, the solution does not need to be generated again. Template based homology modelling is the most common type of structural prediction method used and typically performed in three main steps which rely heavily on dynamic programming. The first task involves identifying a suitable template structure to be used to model the target sequence. This stage is generally performed using

a dynamic programming, sequence alignment algorithm to search within a library of sequences for proteins with structures that have been solved experimentally. It is important to find a structure within the available databases which represents a close evolutionary link; ideally a family member, or at least the most closely related gene available. This is not trivial though and can be problematic for long sequences. Alternatively, a manual search method can also be carried out, identifying genes by looking through free online databases, namely the Protein Data Bank.

4.2.4 Protein homology modelling

The alignment algorithms used in popular programmes such as BLAST (Basic Local Alignment Search Tool) are commonly used by bioinformaticians and incorporated into automated sequence identification pipelines. Other programmes such as CLUSTALW, PRANK, SAling and MUSCLE (for multiple alignments), each use different scoring functions to identify sequence homology and can be substituted in favour for one another. Some may be better for identifying phylogenetic similarities, while others better for predicting shared structural or functional properties. In principal though, each alignment programme identifies positions to map residue pairs which are homologous and generates a framework to build co-ordinates for specific regions of the target structure. To predict the most probable positions, a scoring matrix such as the BLOSUM62 matrix (Supplementary Table 7.1) is used by the sequence alignment programme to give the best overall alignment for matching amino acid residues pairs. It also takes into consideration the physiochemical properties and significance of those residues in proteins. For example, two aligned tryptophans give the highest individual score of 11 for BLOSUM62 because of its unique properties and tendency to be well conserved, whereas a matching valine-pair scores 4, as more alternative residue could perform the same role.

A sequence alignment job may be one against many, by consecutively aligning the target to a library in single jobs, which is typically very fast to compute, or by multiple sequence alignment (MSA) which attempts to make the best alignment for two or more sequence. MSA tools are useful for distinguishing conserved regions in families of proteins and can also be utilised to infer groupings between homologous proteins, evolutionary lineages, and insertion or deletion events. The second step involves using the aligned residues to build a partially complete model, based on the alignment generated in the first step. α-carbon atoms of the backbone are first mapped onto the template. Extended loops or poorly aligned sections are built to avoid steric clashes while maintaining the original structure. Finally, residue side chains are positioned using rotamer libraries and the last step often involves energy assessment and minimisation to identify favourable conformations

[124]. Multiple decoy models are assessed, and a set of the best candidates are generated printed to PDB format with XYZ co-ordinate data.

Although the general steps are largely the same, each modelling software uses unique protocols to generate the model. As such, the models generated could have subtle but significant difference, especially for a large protein such as BamA. I-TASSER is a popular, high-ranking server which uses a method known as iterative threading to generative fragments of the target protein using a diverse library of existing templates [125]. These short fragments which represent the most likely fold for that sequence, according to known structures, are then threading onto a suitable template in an iterative manner until a complete model is generated.

The approach used by MODELLER uses an alternative method which is focused on satisfaction of spatial restraints generated from a library of crystal structures. Essentially, MODELLER works by fitting the sequence to a suitable template identified through homology, and then optimises the structure based on spatial constraints generated from a selection of crystal structures [126]. The spatial restraint parameters used by MODELLER were determined from soluble proteins and for this reason are not ideal for modelling membrane proteins. Nonetheless, previous studies have found that it MODELLER is more accurate in predicted the structure of known membrane protein homologues against other methods such as the SWISS-PROT [127].

PHYRE2 uses a novel approach to template identification by generating a profile using hidden Markov models. This method, known as remote homology detection, incorporates a COMPLETE wide algorithm for structural homology detection using HH-BLITS, an advanced multiple sequence alignment software, and regenerates models based on the most closely related structures available. Each method has its own advantages and disadvantages, and these can be difficult to assess due to the different scoring function used in each case.

4.2.5 Scoring and template fitting

Templates used in protein homology modelling are usually crystal structures which have been determined experimentally. X-Ray crystallography diffraction is widely used to determine the atomic structure of molecules arranged in a crystal lattice. It works on the principle that the atoms of a molecules in a crystalline state diffract an incident beam of X-rays at a specific angle and direction (Chapter 3). It is often applied to solve the structure of many kinds of materials and is a

favoured technique for determining the structure of large, complex biomolecules at high resolution. A crystal structure template is then used to generate a model for a homologues protein sequence and their similarity can be compared be measurement the root-mean square deviation between each structure:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N_{atoms}} d_i^2}{N_{atoms}}} \tag{4.1}$$

where N is the number of distances measured between aligned atoms and d_i is the distance between the *i*th atom of each conformation [127].

Template modelling score (TM-score) is an alternative measurement used to assess the accuracy of fitting, and hence the quality of the model, against a template [128]. The scoring function is weighted by length to eliminate protein size dependence:

$$TM\ score = \text{Max}\,[\frac{1}{L_N}\sum_{1}^{L_T}\frac{1}{1+(\frac{d_i}{d_0})^2}] \tag{4.1}$$

where L_N is length of the native structure, L_T is the length of the aligned residues and d_i is distance between pairs of aligned residues and d_o is used to normalise the difference.

Mechanical analysis is also becoming an increasingly popular method in computational structural biology for model assessment [129]. A protein structure can easily be modelled fundamentally as a mass-spring system from a simple mechanical stand point, where each mass lies across the folded polypeptide chain. Alternative mechanical methods for assessment include molecular mechanics simulation, principal component analysis and normal mode analysis. Normal mode analysis (NMA) is powerful tool for investigating large-scale motions in biomolecular structures. A normal mode is a pattern of motion observed within an oscillating many-body system where components within the system move with correlating frequencies and with a fixed phase relationship. Free motion described by the normal modes can takes place at these fixed frequencies. In terms of protein molecules this is understood by regarding the structure not in terms of single fixed positions, but

instead as an ensemble of many positions which occur over fixed durations of time. What is often referred to as the "native" state of the protein should instead be represented as an ensemble of microstates [130]. These states retain the overall fold of the protein which maintain the secondary structure components.

4.3 Results

*Cj*BamA models were generated using the I-TASSER and PHYRE2 servers running versions 4.4 and versions 2.0, respectively. MODELLER (version 9.14) was run locally on a Debian x86_64 platform using custom Python scripts (available on request). In each case, the programmes generated topologically equivalent models, with all predicted features of BamA including a 16-stranded β-barrel domain and five well defined POTRA domains (Figure 4.1). In each case, the programme identified the *N. gonorrhoea* 4K3B crystal structure as the primary template despite the availability of other models (Table 4.1).

Critical analyses were carried out using the coverage and RMS results from the *align* function in PyMOL. The coverage indicates how many backbone atoms above than the resulting RMSD were excluded during progressive cycles of alignment. RMSD is the magnitude of atomic displacement from the template (Equation 4.1). Hence, a high coverage with low RMSD is favourable and indicates a high-quality homology model. Conversely, a low coverage and high RMSD indicates that the modelling programme could not satisfy the same spatial constraints of the template structure for the target sequence. The results show that the I-TASSER structure has the lowest RMS measurement, ranking highest with the closest fit to the template, however, less α-carbon atoms were fitted in the overall alignment compared to the MODELLER fit. The RMS score for those atoms which did align in the model produced by the I-TASSER is very low and still account for over 70% of the entire backbone. 86% of backbone α-carbons which were aligned in the MODELLER structure also score with a low average RMS measurement of 1.471 Å. The structure generated by PHYRE2 has the highest coverage and was assigned a confidence level of 100% by the automated scoring routine, despite having the highest overall root-mean square deviation from the 4K3B template. This may be a result of the PHYRE2 method for building through remote homology detection, incorporating other more preferential templates in some regions and hence leading to a worse fit against the standard model used here.

Table 4.1 Structural models of BAM proteins from the PDB

PDB Code	*Description*	*Release Date*	*Method*	*Resolution*	*Residue Count*
5LJO	BamABCDE *E. coli*	10/12/2016	cryoEM	4.9 Å	1624
5D0O	BamABCDE *E. coli*	03/9/2016	X-ray Diffraction	2.9 Å	1914
5D0Q	BamACDE *E. coli*	03/9/2016	X-ray Diffraction	3.5 Å	3044
5AYW	BamABCDE *E. coli*	2/24/2016	X-ray Diffraction	3.56 Å	1556
5EKQ	BamACDE *E. coli*	1/20/2016	X-ray Diffraction	3.39 Å	1430
4C4V	BamAΔP1234 *E. coli*	4/23/2014	X-ray Diffraction	3.0 Å	931
4N75	BamAΔP12345 *E. coli*	4/16/2014	X-ray Diffraction	2.6 Å	770
4K3B	BamA *N. gonorrhoeae*	9/4/2013	X-ray Diffraction	3.2 Å	792
4K3C	BamAΔP123 *H. ducreyi*	9/4/2013	X-ray Diffraction	2.91 Å	532

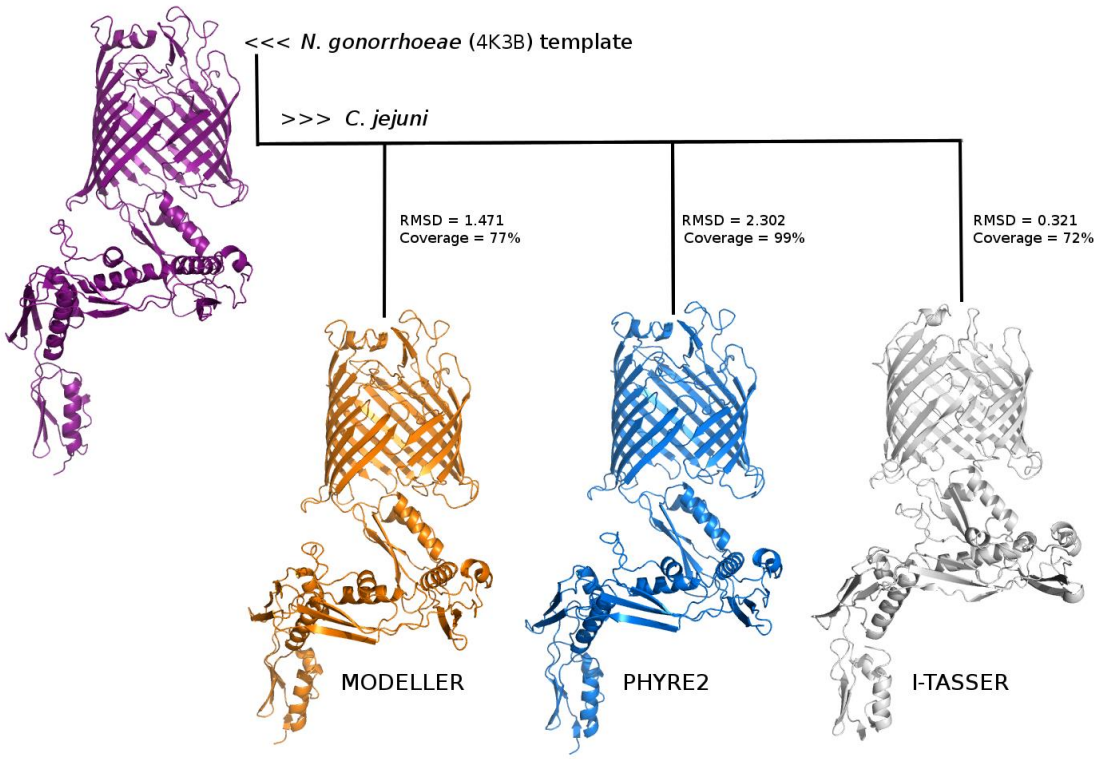


Figure 4.1 Three models of *C. jejuni* BamA are generated using MODELLER, PHYRE2 and I-TASSER. Structural predictions are conserved, with a five-unit POTRA domains and 16-stranded β-barrel all with comparatively low RMSD measurements.

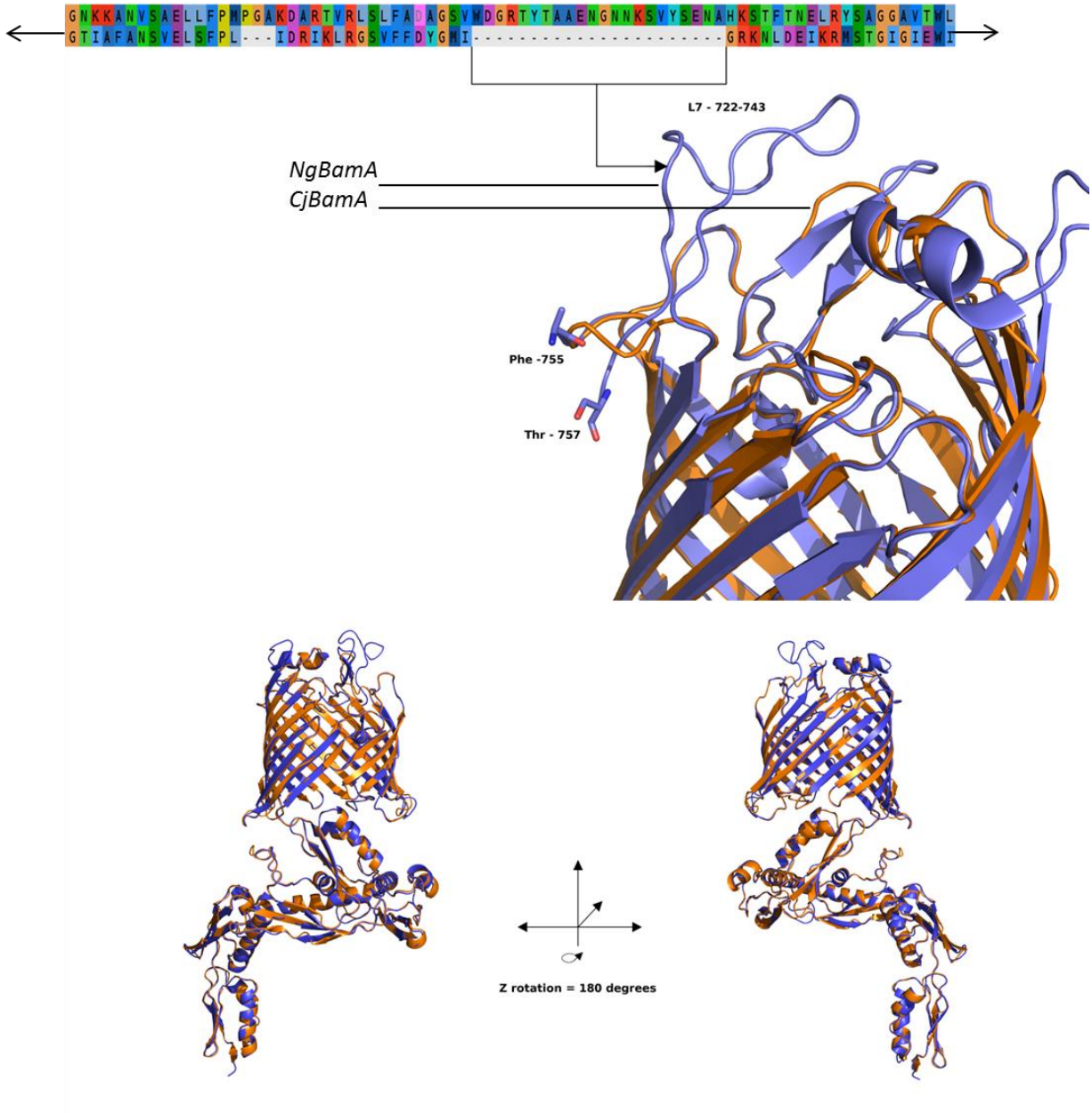


Figure 4.2 Structural alignment of *Cj*BamA model and template 4k3b. Top-right cell shows how the alignment of two homologous sequences inserts a gap where the loop is not present in a *Cj*BamA model generated by I-TASSER (https:// http://zhanglab.ccmb.med.umich.edu/).

The template structure (PDB = 4k3b, purple) aligns well but for larger proteins there is a higher probability of errors in template-based modelling, which is subject to inaccuracies if the initial alignment is poor. These results suggest that the I-TASSER function could be more suited for accurate fitting to a template with more precise refinement in certain regions, whereas MODELLER, while more relaxed about the placement of atoms, could give a better overall global alignment.

Furthermore, the sequence of *Ng*BamA used in alignments featured an extended L3 loop, which was not found in the *Cj*BamA sequence or models. This region was also not resolved by XRCD in the 4k3b model indicating that it did not assume a regular fold in the crystal and could be highly dynamic in nature. Overall though, this region is short enough to not significantly interfere with the alignment.

TM-scores were made using the Zhang lab TM-align server for pair-wise structural comparison between generated models and the 4K3B template (Table 3.2). A TM-score is between 0-1, with 1 being a perfect match. A score below 0.2 is expected for unrelated proteins, whereas above 0.5 the proteins should feature the same overall fold architecture. In all models generated by each method, significant TM-scores were calculated above 0.8. The PHYRE2 and I-TASSER models though were higher than the scores for the MODELLER structure, at 0.922 and 0.925 respectively. However, the PHYRE2 structure RMSD was higher but with a greater coverage, suggesting that the I-TASSER model was more refined overall.

Table 3.2 RMSD measurements coverage and TM-scores against 4k3b template.

Program	*File I.D*	*RMSD*	*TM Score*	*Coverage (%)*
MODELLER	*Cj*BamA_MOD	1.471	0.885	77
PHYRE2	*Cj*BamA_P2	2.302	0.922	99
I-TASSER	*Cj*BamA_IT	0.321	0.925	72

4.3.1 Co-evolutionary Signal Detection

To further understand the structural relationship with the β-barrel assembly machine proteins in Gram-negative bacteria, co-evolutionary signal detection was applied to all available sequences of the BamA family. Firstly, all sequences for known BamA proteins and others inferred by homology were collected from the UniProt database. From over 12,000 sequences, redundant identical entries were removed, and N-terminal signal peptides manually cropped. The non-redundant sequences were then filtered based on a size cut-off between 700 and 900 amino acids in length. The remaining 4,718 sequences were aligned using MUSCLE to generate a multiple sequence alignment and this alignment was further refined using HHblits as part of the GREMLIN server's procedure and filtered to remove sequences with less than 75% coverage and regions with over 75% gaps. The resulting 1,904 sequences were used to detect co-evolutionary signals with a corresponding number of sequences to length ratio of 2.7. This value is used to determine if sufficient sequences are available for assessment, as larger proteins require more unique sequences to achieve the same reliability due to a greater number of potentially insignificant or noisy signals generated from uncorrelated residues.

Analysis of the resulting contact map, which is formed from a prediction of distance between co-evolving residue pairs, shows a characteristic pattern displayed in five POTRA sub-domain regions and sixteen strands associated with the β-barrel (Figure 4.3). Remarkably, when overlaid against a contact map generated from crystallographic data, the backbone residues predicted align almost perfectly across the entire structure. The positions of residue interactions indicate an essential conservation in structure which has been retained over the evolution of BamA sequences, further highlight the value of bioinformatics tools in exploring the relationship between protein sequence, structure, and function.

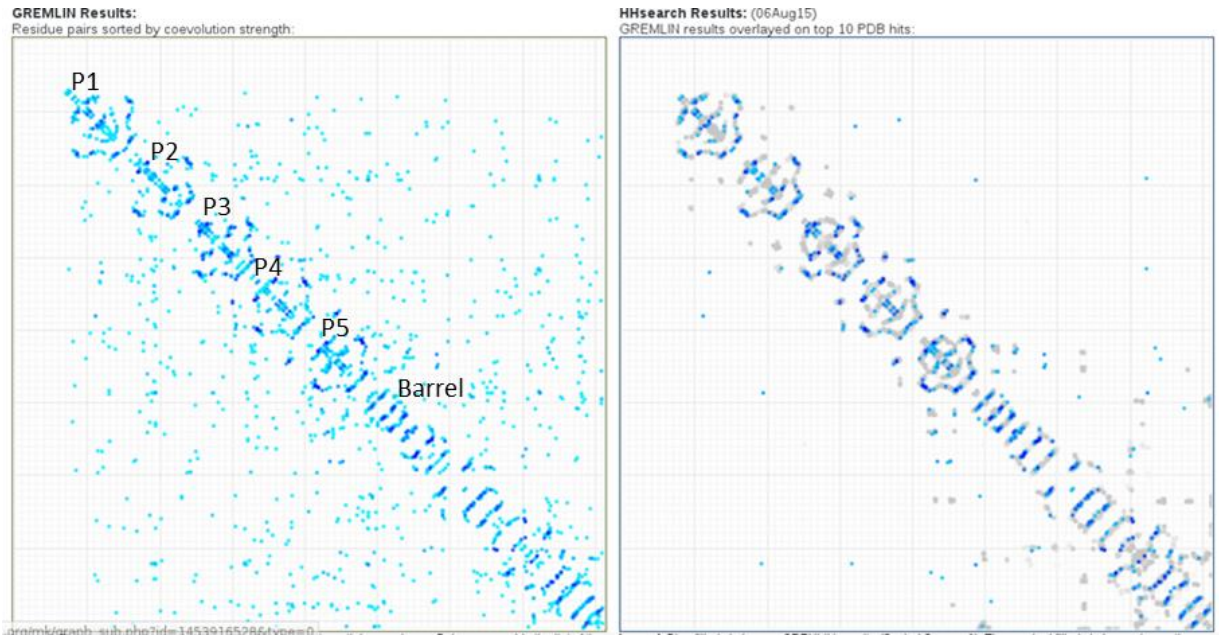


Figure 4.3 Contact map prediction of BamA by co-evolutionary signal detection. The Baker Labs GREMLIN server (http://gremlin.bakerlab.org/) was used to predict co-evolving and conserved residue pair. Contact points are predicted from spatial restraints and generate a two-contact map (left) which shows regular patterns for each POTRA subdomain, and 16 β-strands. The map is overlaid on a contact map generated from crystal data of *Ng*BamA (right) showing the accuracy of this method when provided with a sufficient number of sequences.

4.3.2 Identification of BamD and structural investigation

A previously unidentified gene, showing similarity to the BamD gene in *C. jejuni* was found recorded as a putative membrane protein in the UniProt database. This accessory BamD lipoprotein of the *C. jejuni* BAM complex was inferred from homology from the complete *Campylobacter jejuni* subsp. jejuni NCTC-11168 genome (KEGG, AL111168), as first sequenced in 2000 [20] and shown to be homologous through a basic sequence alignment and comparison method. Firstly, the genome was divided into smaller, manageable sequences and saved as separate files for parsing followed by translation in both directions and in all frames using the EMBL Six-Pack tool [131]. The gene was then identified by aligning a set of known BamD primary sequences from related ε-proteobacteria against all resulting open reading frames translated from the NCTC-11168 genome using BLAST. A single result, aligned with the Campylobacter organism *Helicobacter pylori* (strain ATCC 700392/26695) yielded the best alignment, with a query coverage of 98%, an expectation value of $7x10^{-42}$ and an identity of 35%. This was taken to be sufficient evidence that the *C. jejuni* gene identified, AL111168.1.6 (ORF14209), as shown below, is the missing BamD component from this organism. Furthermore, this sequence aligns well against a putative lipoprotein labelled comL of *Campylobacter concisus* (strain 13826) (Entry: A7ZEJ2) which is now also recognised as part of the BamD family. The sequence of *C. jejuni* NCTC-1168 BamD is as follows:

The translated sequence of the open-reading predicted to be C. jejuni BamD is as follows:
*AL111168.1.6.ORF14209 Translation of AL111168.1 in frame 6, ORF 14209, threshold 1, 216aa:*MKKGIFLLVFLSVFFSACSTKNDEGLYNLSASEWYKQIIKDLQDKDLEKADDHYNGMA
SEHVADPLLETTLIILAQAHMDEEEYKLAEFYLDEYNKKFGNSRNADYIRYLKIKAKFDAFAV
PNRNQALMLESQKEIDTFLKDYPYTEYEPLVQTMLTKFNLAVFYLNSTIENLYQRIGHDESAQ
IYKQRLQESEFYQQSIIKPELPWYRSIFERF

Despite the strong evidence indicating homology between the *H. pylori* and *C. jejuni* sequences, the automated UniProt pipeline had not previously detected this. The gene labelling was subsequently updated which then resulted in a number of other putative lipoproteins in *Campylobacter* species to be found. Although it was surprising that the gene had not been recognised previously using the automated methods, the existence of BamD in *C. jejuni* was not completely unexpected as it has been shown to be essential in most other organisms studied. A similar procedure was carried out in an attempt to find other BamB, BamC or BamE homologues, but the search results did not yield any strong positive results.

Reports of other BamD protein structures show that it has a consensus tetratricopeptide repeat, a regular fold consisting of ten α-helical stretches which turn back on one another in a winding, serpentine-like manner. Using the automated server-based bioinformatics tool, PSI-PRED predicted the secondary structure based on the primary sequence of this recently identified protein. PSI-PRED uses sequence patterns extracted from known protein structures and folds to determine the most probable secondary structure for a given sequence. In this case, the most probable secondary structure of the *C. jejuni* BamD is shown with the expected ten α-helices dispersed along the length of the sequence (Figure 4.4).

Homology models of BamD were generated by the same method to those for BamA, testing the same three programs, MODELLER, I-TASSER and PHYRE2. Table 3.2 shows the RMSD measurements made using the PyMOL *align* function against the BamD structure which was extracted from the *E. coli* BAM complex crystal structure (PDB code = 5D0O). All three models have entirely α-helical secondary structures, arranged in a tetratricopeptide repeat formation. The models generated using I-TASSER and PHYRE2 have well defined α-helix through these repeats as shown in Figure 4.5. In contrast, the model generated using MODELLER has less well defined secondary structure content compared to those from I-TASSER and PHYRE2. This appears to be caused by non-native twists in the α-helix and indicate a defective structure. Furthermore, these structures are used for molecular dynamics simulations to test their structural rigidity. The results show that the MODELLER structure is in fact less stable than its I-TASSER and PHYRE2 counterparts as predicted (please see Chapter 6).

4.4 Discussion

The detection of protein homology infers a shared evolutionary origin in the form of a common ancestor. However, sequences can vary significantly, and it has been recognised that proteins can have relatively low identity yet are able to maintain the same function. In this case, the Omp85 superfamily of proteins has presented an interesting case for studying the relationships between sequence, structure, and function. Namely, BamA modelling must account for both transmembrane and soluble domains, which are unique but have common folds found in nature.

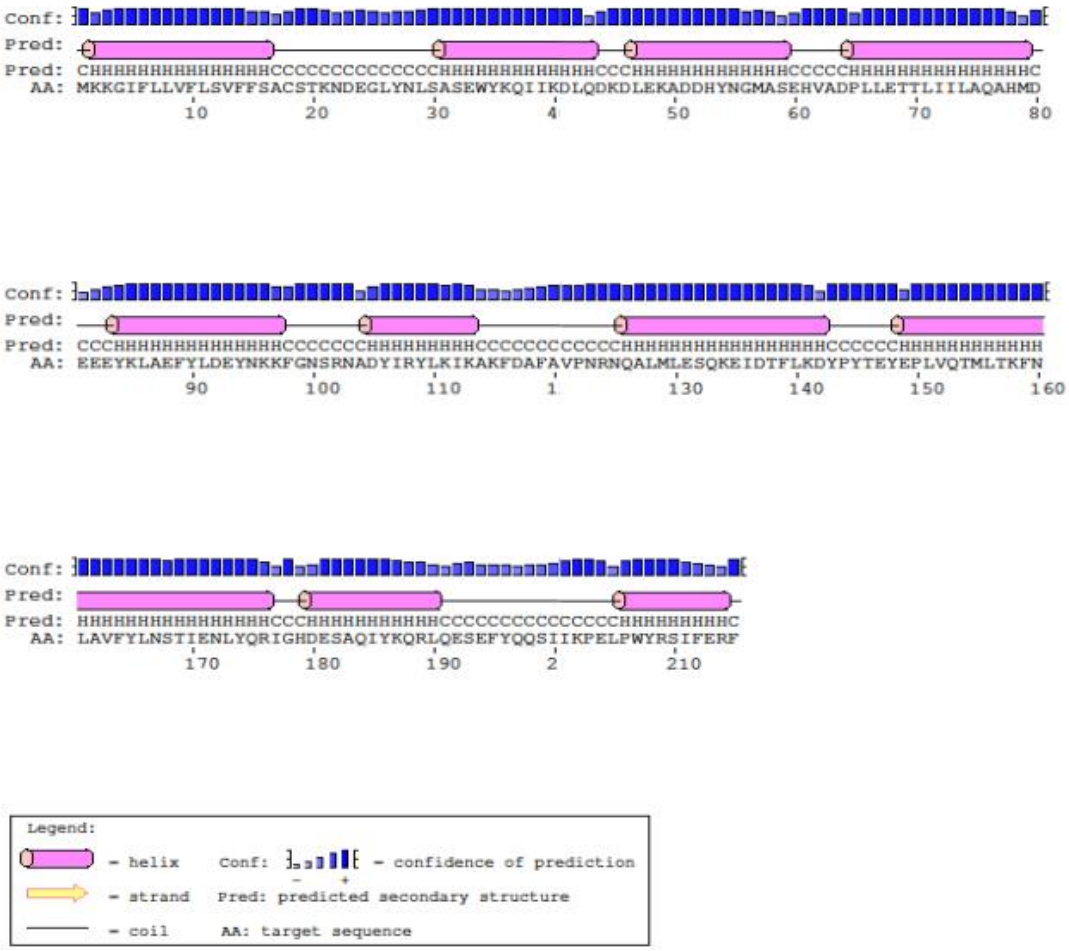


Figure 4.4 Secondary structure prediction of putative *C. jejuni* BamD using PSI-PRED. BamD protein structures have previously been solved showing a tetratricopeptide repeat with ten α-helical subunits. A prediction of the secondary structure motifs based on the primary sequence of BamD gene of *C. jejuni*, recognised through homology, agrees with the known structures of other BamD proteins, solved through XRCD. Ten α-helical motifs are predicted with a high-level of confidence (the first residues 1-16 are associated with the signal peptide for inner membrane translocation).

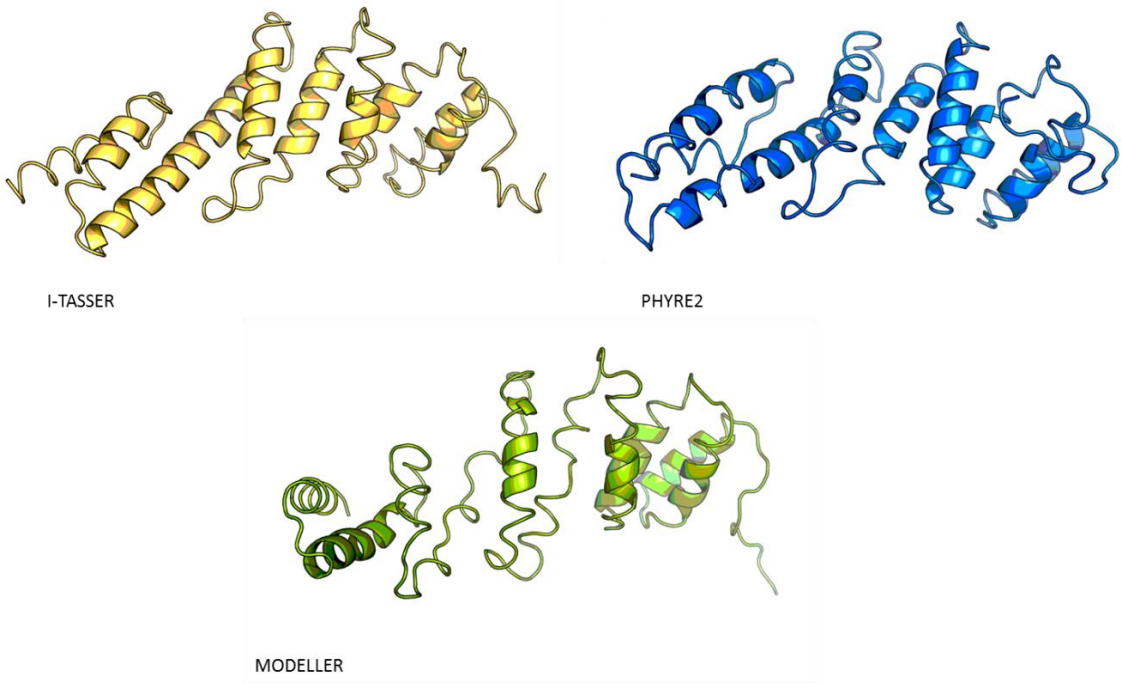


Figure 4.5 Models of *Cj*BamD genereated using MODELLER, PHYRE2, and I-TASSER. Loss of well defined secondary structure is seen in the MODELLER structure, compared to the the I-TASSER and PHYRE2 structural predictions which have α-helical units ordered in a tetratricopeptide repeat fold.

Table 4.2 Assessment of BamD models

Program	*File I.D*	*RMSD*	*TM Score*
MODELLER	*Cj*BamD_MOD	1.01	0.915
PHYRE2	*Cj*BamD_P2	1.83	0.891
I-TASSER	*Cj*BamD_IT	1.16	0.916

4.4.1 The Protein Folding Problem

The number of theoretical conformational poses possible for a protein are massive and as such predicting the correct native fold is challenging. The current understanding of β-barrel assembly in the outer membrane is not lacking though for availability of structural information. To date, there are over twenty entries in the RCSB Protein Data Bank of BamA and accessory BAM complex components alongside several additional Omp85 structural data solved including the *B. pertussis* FhaC protein. Crystal and NMR structures are indispensable resources which need in depth investigation to understand how this group of proteins functions but by determining characteristic structural features which are shared amongst this family, it is possible to learn how sequence information preservation laws have managed to steer evolution processes for this diverse set of molecules.

It is evident that conserved interactions between spatially related regions maintain structural integrity through residue matching, in kind or by co-evolving residue types and interaction swapping. More specifically, two or more residues in close contact may concurrently mutate in an acceptable manner, for example charged residues in a salt bridge or non-polar residues in hydrophobic interaction regions. However, determining how these processes occur with respect to the function of dynamic protein machinery adds another level of complexity to the problem. Alterations made through conformational changes during global structural rearrangements present new time-dependant states in which residues interact with multiple sites, specific to intrinsic protein dynamics. This has already been seen in the differences between two crystal structures of BamA proteins, the 4K3B and 4K3C structures which show two unique conformations of the POTRA domain.

Nonetheless, the generation of homology models is still beneficial and convenient method for inferring dynamics. From the models generated here, a number of shared structural properties relating to function are found. Primarily, the presence of the reduced hydrophobic region between the first and last strand of the β-barrel along with the hook-like L6 loop which resides in the pore have been identified.

These findings highlight the value of incorporating computation methods in structural studies. Ultimately, the question of how structural properties can be related back to biological function can be addressed on some level, without the need for additional high-end experimental techniques

which are both expensive and time-consuming. Of course, this is dependent on the availability of suitable homologous templates and structural data and is not always feasible for every target molecule. Recently though, it has become increasingly apparent that the study of molecular evolution and biophysics need to work more closely together to solve problems relating to protein structure and function [27]. One problem is that phylogenetic inference methods, namely those for establishing evolutionary distances between species, neglect important physically relevant site-site correlations and hydrogen bonding networks which occur between neighbouring amino acids. The assumption that sites evolve independently simplifies the models of phylogenetic likelihood to speed up calculations but does not account for correlated substitutions which must occur to maintain thermodynamically stable, functioning structures.

The current rate substitution matrices - designed to assign the most probable substitutions likely to occur during individual changes in residues - form the foundation of most of the programmes used to study evolutionary links in proteins. They suitably capture physiochemical properties shared amongst amino acids which could be substituted with one another, for example, polar residues replaced for other like polar-residues or hydrophobic groups switching with other similarly charges residues. These methods though generally take no information relating to structure and therefore treat each residue in a similar manner. As such, this simplistic approach to phylogenetic inference can be misguided as it fails to account for the local environment in which the amino acid might reside.

Even if these shortcomings described are overcome, further complications arise when investigating folding pathways. Mutations which are accepted in the folded state may impact on folding pathways, having a dramatic consequence on the final state of the protein. Furthermore, intrinsic protein dynamics could lead to substitutions which are not occurring in a rigid body system. As such, it is not possible to determine the exact position of a single residue in relation to the rest of the structure at any one time. Although these problems can be addressed using more sophisticated molecular dynamics simulations, as shown later, methods for accurately assigning relative co-ordinates in a static protein structure are inherently flawed because of their dynamic nature. It is also due to the flaws and limitations of the crystal structures used as starting information in modelling jobs meaning that despite being the most efficient method for predicting structures, homology modelling can never completely reliable.

Here, the *C. jejuni* BamA sequence has been subjected to a number of other bioinformatics and structural assessment tools which have not previously been reported, including structural prediction using a set of template-based homology modelling programmes. We see that despite a low sequence identity below 25%, the predicted structure adopted agrees with expectations. Critical assessment of models generated against crystal templates show that the programmes used can lead to subtle but potentially functionally limiting changes in the three-dimensional structure. Furthermore, a refined multiple sequence alignment of over 1,200 BamA proteins used to predict contact points through co-evolving residues also provides results that are in strong agreement with both the predicted model and the experimentally determined structures.

Furthermore, there is an extensive collection of sequences available for BamA proteins and other Omp85 superfamily members in bacteria as well as in other Eukaryotic organisms. The BamA family of proteins has presented a practical example of this with some family members sharing less than 20% sequence identity yet thought to impart the same action [39].

Detecting homology and subsequently generating models has been a valuable exercise for investigated a target structure and studying the structural evolution of biological macromolecules. In conclusion, the models generated for *C. jejuni* BamA and BamD all retain the same structural features as seen in high-resolution crystal structures of BAM proteins solved experimentally. Analysis of secondary structure, RMSD measurements and TM-scores gives more detail of the model quality against the templates used. Investigating models generated using other templates could increase confidence and provide further insight into the structural relationships between BamA and other Omp85 superfamily members.

Chapter 5. Secondary structure prediction, estimation and thermolytic degradation by synchrotron radiation circular dichroism

5.1 Introduction

High resolution structure determination of BAM complex proteins is a critical step towards solving the mechanism of β-barrel assembly. It has previously been achieved through X-ray crystallography, nuclear magnetic resonance, and electron microscopy - proving a structural basis to model the dynamic process of assembly – although a clear mechanistic solution has not yet been determined. A functional understanding of complex biochemical processes such as β-barrel assembly is often limited by technical difficulties in preparing samples and an inability in performing time-resolved experiments, making it difficult to translate structural data into functional dynamic models. To gain further structural understanding in the absence of high-resolution data, circular dichroism is proposed to provide insight into the secondary structure of *Cj*BamA and evaluate theoretical models generated earlier. As shown previously, successful expression and purification of *Cj*BamA was confirmed using SDS-PAGE and Western blotting and yet preliminary crystallisation trials were aimed to further characterise this protein were inconclusive. For this reason, the three-dimensional structures generated using homology modelling tools earlier, supporting the initial hypothesis that the structure of BamA is predominantly conserved amongst Gram-negative bacteria, are compared against secondary structures determined by *in vitro* methods. Circular dichroism results are in agreement with the secondary structure analysis of the models, both sharing a dominant β-sheet fraction, in line with the homology models generated. Furthermore, thermolytic degradation is used to investigate the stabilising effect of two different detergent buffers used to solubilise BamA.

5.2 Background and Theory

Circular dichroism (CD) is a powerful spectroscopic technique used for investigating protein folding and secondary structure composition. It is also a useful tool for studying structural stability, protein-ligand interaction, and optimising buffer conditions by radiolytic and thermolytic degradation. One important feature is the ability to analyse membrane, as well as soluble proteins. Circular dichroism works by measuring the differential absorption of circularly polarized light -

the difference between left-hand (LH-CPL) and right-hand (RH-CPL) circularly polarised light – at different wavelengths which is used to infer the relative orientation of the peptide backbone and obtain low-resolution structural information of proteins in solution. Furthermore, circular dichroism is commonly used to investigate folding and secondary structure components of proteins by time-resolved estimation of α-helix, β-sheet and turn fractions [132]. Unlike XRCD, these components can be monitored in a time-dependant manner at physiologically relevant conditions and in the presence of binding partners or at different atmospheric conditions, for example temperature or pressure.

The difference in absorbance between left-hand and right-hand circularly polarised light is used to determine the molar ellipticity, which is weighted value used to determine secondary structure. Firstly, the difference in absorbance is described by the following equation:

$$\Delta A = A_{LH} - A_{RH} \tag{5.1}$$

where ΔA is the difference in absorbance between left hand (A_{LH}) and right hand (A_{RH}) circularly polarised light. By applying Beer-Lamberts law, the difference in absorbance is described as follows:

$$\Delta A = (\varepsilon_{LH} - \varepsilon_{RH})Cl \tag{5.2}$$

where ε_{LH} and ε_{RH} are molar extinction coefficients, C is molar protein concentration and l is path length in cm. The molar ellipticity is then defined as:

$$\Delta\varepsilon = \varepsilon_{LH} - \varepsilon_{RH} \tag{5.3}$$

where $\Delta\varepsilon$ is a per residue, molar absorption unit of circular dichroism ($M^{-1}cm^{-1}$) otherwise known as molar circular dichroism. Mean residue ellipticity, which is most commonly used to express the circular dichroism of the substance and determine secondary structure, is expressed using the relationship:

$$[\theta] = 3298.2\,\Delta\varepsilon \tag{5.4}$$

5.2.1 Secondary Structure Estimation of Protein

Secondary structure determination of proteins is used to characterise the folded profile of a purified sample or of a structure solved at atomic resolution. Two mathematical models which are applied to estimate the secondary structure content for *in vitro* protein samples and PDB structures are described. As shown below (Figure 5.1), the CD spectra is collected in the far-UV range between 180-260 nm can be used to determine the contribution of α-helical, β-strand and disordered, also referred to as random coil, regions in a folded protein. In this spectral region, the peptide bonds in the protein backbone acts as the chromophores which have characteristic absorption profiles when the amino acids are found in a structured region. Residues found in α-helical folds give rise to negative spectral bands at around 208 and 220 nm and a positive band near 193 nm, whereas residues in β-sheet conformations contribute negative and positive bands at 218 and 195 nm, respectively. To extract secondary structure content, the spectra can be simply treated as a linear combination of all contributing secondary structure types. However, this method is generally not useful, or accurate, as there is no reference set which describes the CD spectra of proteins composed entirely of only a single secondary structure type [133]. Because of this, a reference set of known structures is used to compare the spectra and assign the fold composition. It is important though that an accurate protein concentration is determined in order to be useful for comparison against the reference set. This is due to a need to incorporate the molarity of solution in the algorithms so that samples with different concentrations can be analysed systematically.

5.2.2 Protein Concentration Determination

To calculate secondary structure using circular dichroism spectroscopy, a known protein concentration of the sample under investigation is needed. Beer-Lamberts Law is used determine the concentration of the protein in solution using the following equation

$$OD = \varepsilon C l \tag{5.5}$$

where OD is optical density, epsilon is molar extinction coefficient, L is the path length (cm) and *C* is the concentration of the protein in solution (M).

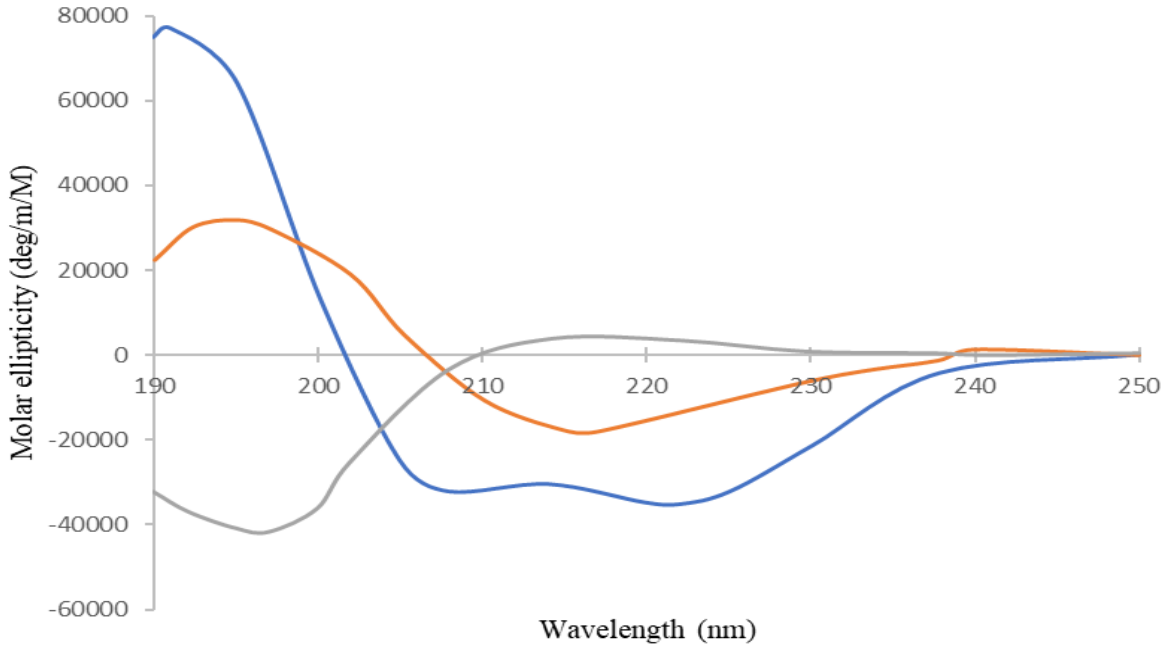


Figure 5.1 Reference spectra for secondary structure estimation using circular dichroism. Purified poly-l-lysine can assume both α-helix (blue) and β-sheet (orange), and undefined (orange) secondary structures. This model is used to provide a reference for deconvoluting the CD spectra of novel proteins [134].

The approximate molar extinction coefficient of the protein is calculated based on the number of tryptophan, tyrosine and cysteine residues by weighting the sum of the absorption coefficient for these three residues at 280 nm is approximated by [135]:

$$\varepsilon = (n\mathrm{W} \times 5500) + (n\mathrm{T} \times 1490) + (n\mathrm{C} \times 125) \quad (5.6)$$

5.2.3 CONTIN/LL

The CONTIN/LL method for protein secondary structure estimation from CD spectra uses a large reference set of soluble, denatured proteins and membrane proteins of known structure. The basic principle used in this method to estimate secondary structure, is that the protein CD spectrum can be expressed as a linear combination of each individual secondary structure component. Therefore, the spectrum is the sum of each component which be extracted using a deconvolution algorithm and reference set [136]. The SMP56 set includes 42 soluble proteins and 14 membrane protein references with secondary structures determined from atomistic models using methods such as DSSP or STRIDE [133]. It is important to obtain CD spectra in the far-UV region as low as 178nm for accurate α-helix and β-strand component estimation [134]. However, the CONTIN/LL method has been shown to perform best over other methods (SELCON, CDSSTR) with a larger reference set for a shorter wavelength range [137] in the 190-260 nm region. Although not entirely accurate, these methods provide a useful and reliable quantitative analysis of secondary structure composition.

5.2.4 DSSP

The *Define Secondary Structure of Proteins* (DSSP) method is a hydrogen-bond estimation algorithm used to assign secondary structure to residue co-ordinates from a PDB structure. Devised in the early 1980s by Kabsch and Sanders [138], and originally written in the Pascal computer language, DSSP has become a standard method for defining secondary structure assignments to solved protein structures and theoretical models. It does this by calculating the hydrogen bond energy between all backbone atoms using an electrostatic Coulombic approximation. It then determines the most likely secondary structure to which each residue belongs, assigning them to one of eight groups belong to three classes: helix (α-helix, 3_{10}-helix, π-helix), strand (β-bridge and bulge) and loop (turns, curved and undefined).

DSSP is advantageous in some regards as it is fast to compute and can be used to accurately define the secondary structures of both experimentally solved and computationally generated model structures. However, it has no *in vitro* application and therefore cannot be applied experimentally but it can be used to assess models and validate findings gathered from CD experiments, making it a useful addition to the CONTIN/LL method.

5.3 Experimental Set-up

Circular dichroism spectra were collected between 180-260 nm. Samples were prepared using 0.1% DDM and 0.5% SB-12 in 50mM Tris-HCl, 200mM NaCl, pH 7.5 and protein concentration was determined using a Nanophotometer at 280nm as described below. Constant temperature measurements were collected at 20°C, taking four repeat scans with an acquisition time of 1 second. Data was processed using the CDApps programme (Diamond, B23) with the CONTIN/LL method and SMP56 reference sets (including 16 membrane proteins of known structure).

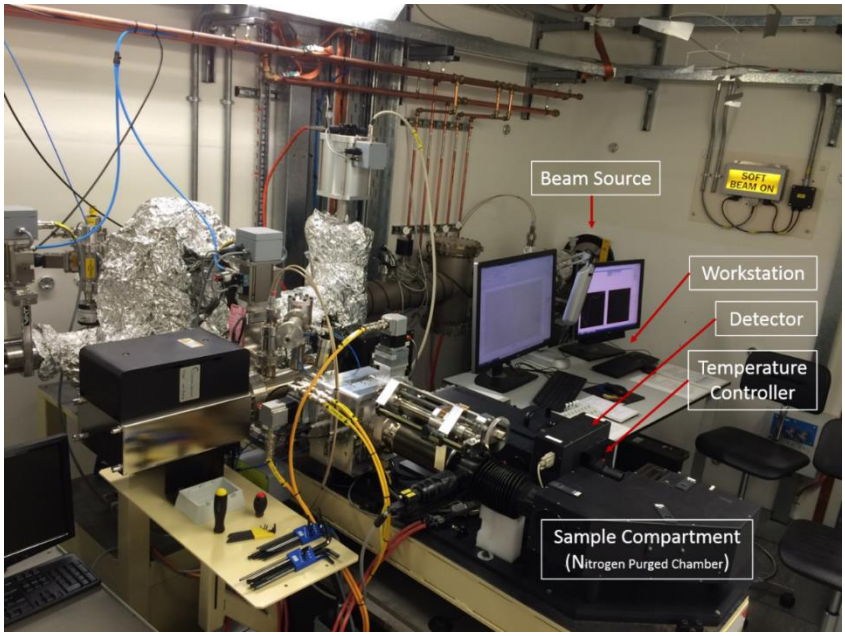


Figure 5.2 Synchrotron radiation circular dichroism instrument rig used to analysis the secondary structure of *Cj*BamA. The storage ring is located, out of picture, to the right - the intensity of light available is much higher to facilitate analyse of samples at lower concentration and with greater accuracy (Diamond Light Source, Harwell, UK).

Total protein concentrations were calculated using a P-Class NanoPhotometer (Implen, P-class) using a molecular weight of 83.27kDa for the full-length protein with TEV protease site and 10x histidine tag. The extinction coefficient calculated as described above was 98,560 (ExPASy, ProtParam).

5.3.1 Thermolytic Degradation

Samples were prepared using standard protocols as previously described in Chapter 2 using 0.03% DDM and 0.5% SB-12 in 50mM Tris-HCl, 200mM NaCl at pH 7.5. Constant temperature measurements were collected at 20°C taking 4 repeats scans with an acquisition time of 1 second. For thermolytic degradation studies, single scans were taken between 20°C and 90°C with an acquisition time of 2 seconds. Reference buffer (Tris-buffered saline with detergent) spectra were acquired using the same set-up. Spectra were processed with CDApps using the CONTIN/LL method and SMP56 reference set. Prior to analysis, spectra were offset to the baseline between 255 and 258 nm and the signal was then restricted to between 190 and 260nm. Transition melting points were calculated at 208, 218 and 220 nm using a sigmoidal Boltzmann fitting procedure.

5.4 Results

In order to investigate the folded profile and confirm earlier predictions about the structure of *C. jejuni* BamA we sought to acquire CD spectra in the far-UV region for this membrane, folded in a suitable membrane mimetic. Circular dichroism was used successfully to determine the fold composition of *C. jejuni* outer membrane BamA preperations in a number of detergent solutions after primary IMAC, namely, SB12 and DDM, as confirmed by SDS-PAGE and Western blot analysis. Protein concentrations in each case were measured above 0.3 mg/mL by absorbance at 280 nm using an Implen P-Class Nanospectrophotmeter. Measurements were made by 1 second acquisitions of clean buffers without protein to determine the absorbance of salts and Tris at the concentrations used. CD analysis was then conducted using repeated scans of the samples in a range of temperatures up to 95°C. Secondary structure analysis at 25°C estimated using the CONTIN/LL method and SMP56 reference set shows a dominant β-sheet content in both 0.1% SB12 and 0.03% DDM solutions, followed by reconstituted protein in DMPC liposomes.

5.4.1 Secondary structure estimation

The fraction of secondary components was estimated using the CONTIN/LL method for both detergent systems. The proportion of β-sheet estimated was almost 10% higher for the *Cj*BamA-

SB12 sample and reduced levels of β-sheet folding in the *Cj*BamA-DDM sample are observed along with a proportional increase in disorded regions. However, the α-helical content was similar in each, below 10% overall. The reconstituted sample in 1,2-dimyristoyl-*sn*-glycero-3-phosphocholine (DMPC) displayed a higher α-helical content. Comparision between the predicted secondary structure content, calculated by DSSP, for the top I-TASSER model generated (Figure 5.4.1) is more closely matched to the higher β-sheet content estimated in SB12, while the fraction of turns is within 5% in both DDM and SB12 solutions. Reconstituted BamA in DMPC with a higher α-content, estimated at 20%, is closer to that expected from the template-model.

Along with comparible α-helical content, the fraction of turns estimated in both DDM and SB12 solutions are almost identical yet the β-sheet content is significantly lower in DDM, which appears to have lost structure to more disordered regions. This result suggests that the strands forming the β-barrel are less well defined within the DDM detergent micelle and that the folding profile is affected by the detergent used to solubilise this domain. Both detergents share an *n*-dodecly hydrocarbon chains but the headgroups differ giving them each unique properties which might more or less preferable to binding with the BamA β-barrel.

The discrepancy between folding profile in DDM and SB12 may indicate that the fold of the BamA protein is attributed to the choice of detergent. However, the shift between the two estimated fold compositions may instead be due to an increased level of aggregation or increased concentration of misfolded protein. It is possible that each sample is comprised of different states of BamA, as well as low levels of contaminants or truncated forms. While each detergent can maintain BamA in its regular conformation with the same topology, the fraction of this species is higher in one detergent suggesting the fraction of properly folded species has shifted due to the nature of the detergent protein interactions.

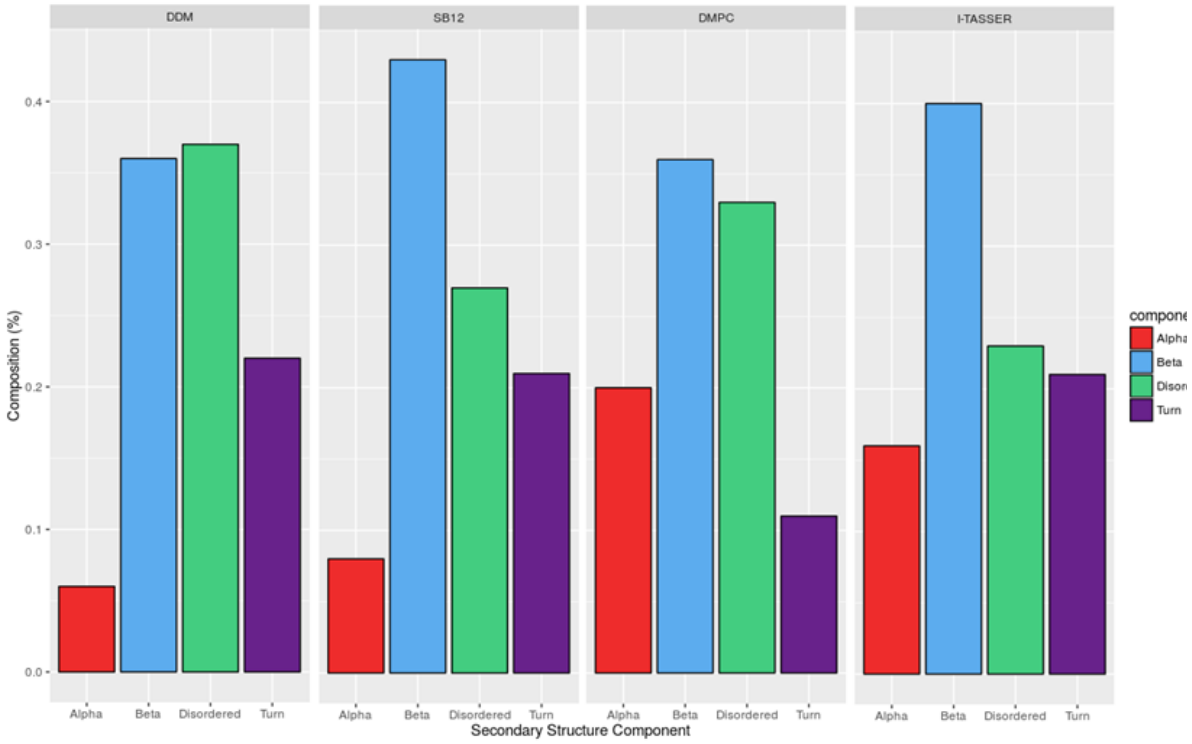


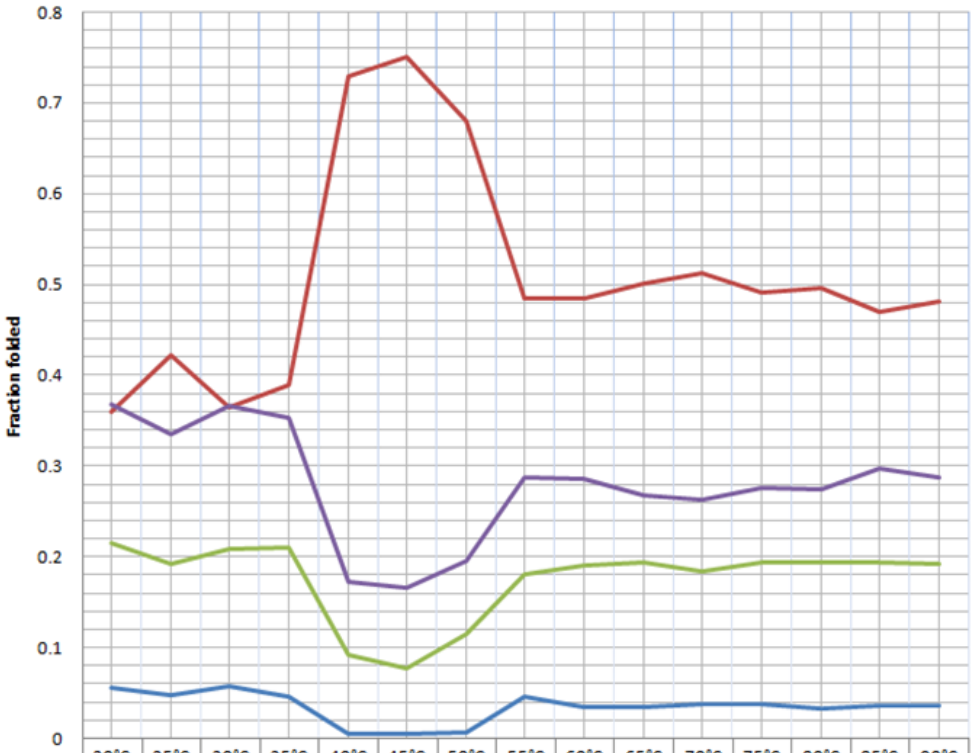
Figure 5.3 Secondary structure estimation of *Cj*BamA. SSE was performed using the CONTIN/LL method with SMP56 reference set for three *Cj*BamA samples at 25°C. The secondary structure of the I-TASSER homology model is shows as calculated using DSSP.

5.4.2 Thermolytic degradation and protein stability assay

Thermal melts were performed between 20 and 95°C using an integration time of 2 seconds with a single scan for each wavelength. Spectra were processed using CDApps (Diamond B23). Secondary structure estimation using the CONTIN/LL method and SMP56 reference set performed using the CD spectra acquired at each temperature point measured in 5°C step intervals from 20 to 90°C for DDM and SB12. Secondary structure estimated at each temperature point (Figure 5.4). The sample in 0.03% (w/v) DDM solution appears to go through three separate phases of folding. The first transition between 35 and 40°C leads a total loss of α-content and a significant increase in estimated β-sheet content. After the protein melts it assumes a conformation largely dominated by β-sheet folds. Similarly, to the DDM micelle system, BamA in SB12 also goes through a transition prior the melting, when it adopts a similar state to that of DDM above the T_m.

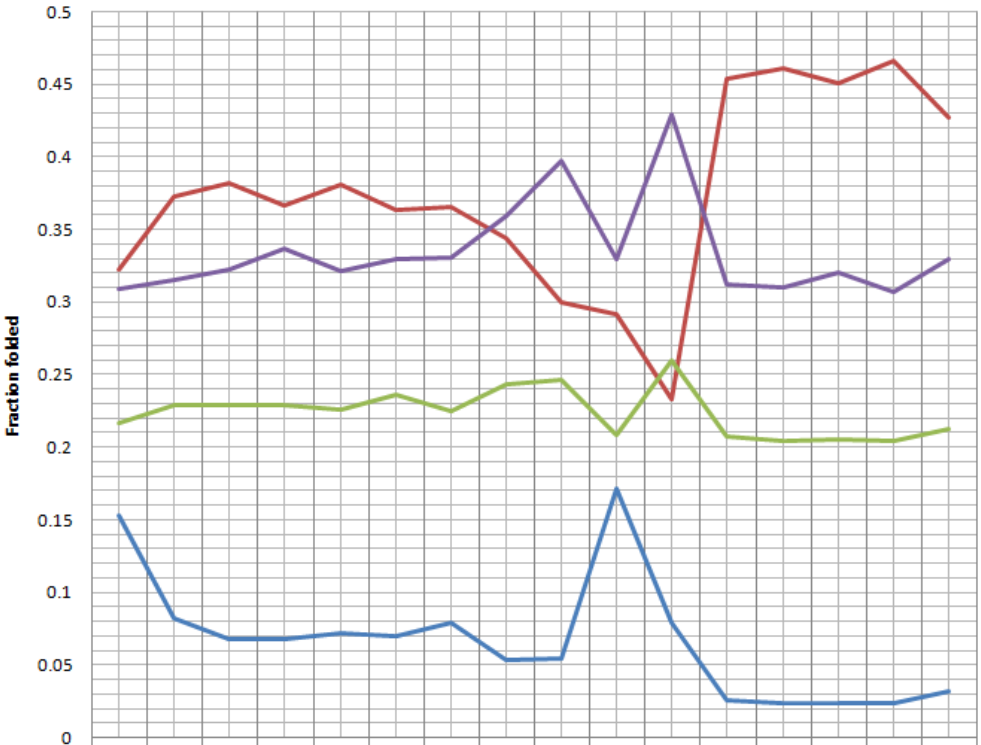
However, unlike in DDM this transition shows a reduction in β-sheet content and an increase in α-helical folds estimated (Figure 5.6). Furthermore, the transition phase before the protein melts begins at around 15°C higher temperature in SB12 than in DDM. Due to the distinct topology of Omp85 proteins, which feature a large transmembrane beta-barrel structure and extended multi-domain soluble POTRA (polypeptide associated transport) region, this 'pre-melt transition phase' may occur as the POTRA regions become highly dynamic and begin to partially unfold as the temperature increases. A Boltzmann fit was used to determine the melting point (Tm) in each case indicating that the transition temperature

These findings were supported using two detergent buffer stability assays assessed by SDS-PAGE. Ultracentrifugation dispersity sedimentation assay was previously used to determine how the protein aggregation affect occurs over time. Furthermore, a heat modifiability assay was used to determine how the protein behaves at elevated temperatures by heating samples up to 65°C before being run on a gel under normal conditions (Figure 5.4). It was that in SB12, the sample behaved normally whereas in DDM, the band was lost and appeared as large aggregated material which entered at the top of the gel. Again, this result indicates that the SB12 protein-detergent micelle complex has increased stability and is more resistant to heat.



	20°C	25°C	30°C	35°C	40°C	45°C	50°C	55°C	60°C	65°C	70°C	75°C	80°C	85°C	90°C
Alpha Helix (Both Types)	0.05	0.04	0.05	0.04	0.00	0.00	0.00	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Beta-strand (Both Types)	0.36	0.42	0.36	0.38	0.72	0.75	0.67	0.48	0.48	0.50	0.51	0.49	0.49	0.47	0.48
T: Turns	0.21	0.19	0.21	0.21	0.09	0.07	0.11	0.18	0.19	0.19	0.18	0.19	0.19	0.19	0.19
U: Unordered	0.36	0.33	0.36	0.35	0.17	0.16	0.19	0.28	0.28	0.26	0.26	0.27	0.27	0.29	0.28

Figure 5.4 BamA-DDM thermal melt folding profile. A solution of BamA prepared in DDM was heated in 5°C intervals every minute and CD acquisitions where. The processed data shows that the composition of protein fold is stable until around 35oC at which point the solution undergoes a transitionary period before returning to stable state at 55°C with a high β-sheet fold content associated with the formation of plaque-like aggregates.



	20°C	25°C	30°C	35°C	40°C	45°C	50°C	55°C	60°C	65°C	70°C	75°C	80°C	85°C	90°C	20°C
Alpha Helix (Both Types)	0.15	0.08	0.06	0.06	0.07	0.07	0.07	0.05	0.05	0.17	0.07	0.02	0.02	0.02	0.02	0.03
Beta-strand (Both Types)	0.32	0.37	0.38	0.36	0.38	0.36	0.36	0.34	0.3	0.29	0.23	0.45	0.46	0.45	0.46	0.42
T: Turns	0.21	0.22	0.22	0.22	0.22	0.23	0.22	0.24	0.24	0.20	0.26	0.20	0.20	0.20	0.20	0.21
U: Unordered	0.30	0.31	0.32	0.33	0.32	0.33	0.33	0.35	0.39	0.33	0.42	0.31	0.31	0.32	0.30	0.33

Figure 5.5 BamA-SB12 thermal melt folding profile. The solution of BamA prepared in SB12 was heated in a similar manner at 5°C intervals every other minute followed by CD acquisition. The solution undergoes a transitionary period between 60°C and 75°C before returning to another stable state with a 45% β-sheet fold content. This transition occurs some 30 – 40°C higher than in a sample prepared in DDM, indicating that the SB12 micelle imparts better stability to the BamA structure.

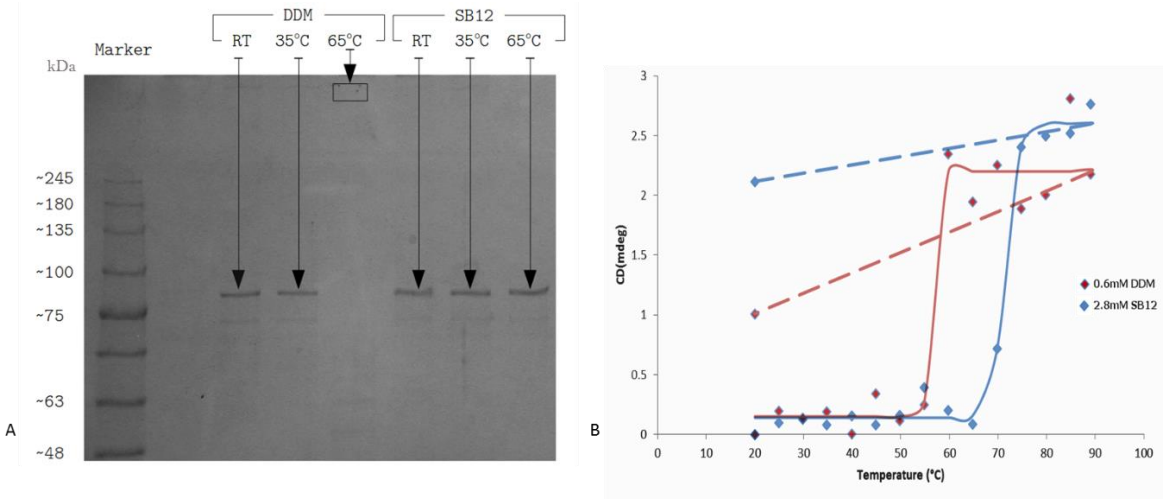


Figure 5.6 Heat modifiability of BamA. The melting point transition is calculated at T_m = 55.08°C for BamA solubilised in 0.6mM DDM-TBS buffer solution. In the same buffer, replacing the non-ionic DDM detergent with 2.8mM of Zwitter-ionic SB-12 in the same buffer gives provides a stabilising effect of over 15°C, with a new transition point measured at T_m = 71.42°C as indicated by thermolytic degradation (B). SDS-PAGE heat modifiability (A) also shows high molecular weight protein aggregates are formed when BamA is heated to 65°C in DDM.

Physiochemical properties of each detergent could explain this effect. These properties are attributed to the detergent chain length, headgroup charge and size, related to measureables including critical micelle concentration and aggregation number. Critical micelle concentration is the concentration at which spontaneous micelle formation occurs, and aggregation number is the number of monomeric detergent units which assemble to form a single micelle. Comparing the CMC and aggregation numbers of each detergent can indicate propensity micelle formation and relative packing geometries. The CMC of DDM is approximately 0.12 - 0.17mM, while that of SB12 is much higher between 3.3 – 3.6mM. Furthermore, the aggregation number of DDM and SB12 are 98 and 55 respectively. This suggests that DDM has a higher propensity to form smaller micelles compared to SB12. However, the folding data suggests that once the SB12 micelle has formed a complex with BamA then it is more resistant to heat.

In general, the trend observed was that the BamA protein would attain a higher level of β-strand component after the transition point. The protein was observed to be stable, with a higher temperature in 0.1% SB12 compared to 0.03% DDM.

5.5 Discussion

Circular dichroism spectra of a *Thermus thermophilus* Omp85 collected using bench-top instruments have previously been reporting indicating a predominantly β-sheet secondary structure [139]. However, no reports were found to provide a detailed secondary structure analysis using synchrotron radiation. The secondary structure content of folded *Cj*BamA protein solubilised in two detergent buffer systems, DDM and SB12, and reconstituted into DMPC lipid membranes are characterised. Thermal melt essays for each sample are performed and compared to results from a heat modifiability and ultracentrifugation dispersity sedimentation (UDS) SDS-PAGE assays. Furthermore, the application of thermolytic degradation, used to study the stability of the protein in various detergent buffers, is thought to be a novel application for tailoring membrane protein preparation steps which could be more widely applied to screen for suitable detergents which can decrease the rate of aggregation and facilitate long term storage which could improve the chance of success for further high-end structural studies. In this section, synchrotron radiation circular dichroism is used to investigate the secondary structure of purified, detergent solubilised samples of BamA and how this changes in response to temperature.

Secondary structure analysis of *Cj*BamA confirmed a dominant β-sheet fraction, attributed to the large 50kDa β-barrel domain. Although the proportion of secondary structure components estimated did not exactly match those calculated by DSSP from the homology model and other XRCD structures, they were still within reasonable limits. Analysis of α-helical content was, however, shown to vary significantly in two cases of detergents used. High salt concentrations were used to maintain stability. However, in solution, NaCl is spectroscopically active below 200nm and low concentrations are preferential to obtain accurate readings in the far-UV signal and for subsequent SSE analysis. Furthermore, Tris acidified with HCl is only recommended at low concentration between 10-20mM [140].

Circular dichroism spectra thermal melt point analysis and SDS-PAGE heat modifiability assays both indicate that SB12-TBS is a more suitable detergent buffer for maintain BamA in a soluble folded form. These findings could be used in conjunction with data indicating the likelihood for membrane proteins to aggregate in each detergent buffer solution or the propensity to crystallise based on the class of target structure predicted. For example, the results shown here suggest that an SB12 buffer enhances the stability the BamA solution in response to elevated temperatures compared to the use of DDM. The strength of intermolecular interactions between the hydrophobic regions of the protein and the detergent tails could be inferred relatively based on the temperature gap between melt points in each case. SB12, for instance is generally less frequently used than DDM in crystallisation trails. This is maybe, perhaps due to weaker forces between detergent chains in the micelle and the protein which allows transmembrane domains to escape the micelle, thereby improving chances of crystal formation.

This approach would be useful to improve systems for rational design and selection of detergent buffers used at every stage of the purification and crystallisation process as different detergents would be preferential at different stages. Whereas a non-ionic, low critical micelle concentration detergents such as DDM may be well suited for crystallisation, Zwitter-ionic and higher CMC detergents such as SB-12 or CHAPS might be more useful for purification, when the protein is subjected to increased shear stress, or for use in long-term storage. However, detergent properties alone are likely to be sufficient to speculate about such applications. Thermolytic degradation assays though provide meaningful results which can be used to quantitatively assess and screen for ideal buffer conditions before expending high material costs on trial and error approaches.

Chapter 6. Molecular dynamic simulations of BAM proteins

6.1 Introduction

Molecular dynamics simulation is a principal tool used to study the dynamics and behaviour of biological macromolecules. It provides an extensive toolbox for probing simulated biomolecular systems on an atomic level by computing the time-dependent behaviour of molecules and provides further insight into their structural and functional properties, complementing crystallography, NMR or other experimental techniques. Recent advances made in computer hardware and algorithm optimisation have made molecular dynamics (MD) simulation programmes much more powerful, and capable of sampling a greater number of conformations for biomolecules over biologically relevant time-scales.

Due to the increased availability of high-performance computing (HPC) facilities, MD simulations have been applied to larger systems to investigate essential dynamics, protein folding pathways and binding partner interactions. In this chapter, studies of full length BamA and BamD models are presented using all-atom (CHARMM36/GROMOS) and course-grained (Martini) simulations run in two MD programmes - GROMACS and NAMD. Analysis of the static *C. jejuni* BamA I-TASSER model structure is performed using normal mode analysis to predict hinge-like regions and display the POTRA domains with enhanced mobility. This observation is then confirmed in short repeat 10 ns all-atom simulations and in a single 111ns simulation. Normal mode, principal component and harmonic analysis are used to visualise correlated dynamics which might be related to function. Finally, the structure of three BamD models are assessed using MD simulation to investigate structural stability.

6.2 Background and Theory

Essential protein dynamics are difficult to determine using normal experimental approaches. The models generated using X-ray crystallography and NMR ensembles provide hints clues but typically several techniques need to be applied in combination to thoroughly investigate how a protein is behaving in real-time, and this is especially true for membrane proteins. As already

shown, despite the wealth of structural information, the assembly mechanism of the BAM complex is still putative. However, since its conception, molecular dynamics has presented an alternative route to investigated protein behaviour, enabling users to explore complex structural rearrangements relating to function in real-time and with absolute atomic resolution. The main advantage is that unlike experimental methods used to investigate protein folding and function, MD simulations provide a continuous trajectory of atomic resolution co-ordinates, known as frames. They can be used to support and complement experimental findings, starting with a structure determined *in vitro*, or they can be used on entirely synthetic systems using models generated from homology software, as described previously in Chapter 4.

However, these simulations are based on force-fields which do not represent an exact model of the forces which govern protein folding. Furthermore, they require considerable computational power to calculate [141]. Early MD simulations were used in the 1990s using the same underlying principles. These simulations were much smaller though, and at the time a short 5 amino acid residue peptide simulated for only a few nanoseconds was considered 'extremely long' simulations [142]. Since then, hardware has improved greatly, and MD simulations are frequently used to study the dynamics of soluble and embedded membrane proteins on the micro-second time scale.

Advantages of MD simulation also include the ability to manipulate starting structures and conditions. For example, molecular dynamics studies of the *N. meningitidis* autotransporter NalP protein were carried out using a truncated model of the crystal structure with the α-helical plug removed [5]. NalP has an autotransporter domain β-barrel located at the C-terminal and a passenger domain including an α-helix that plugs the barrel and which is cleaved and transported out of the cell. Through molecular dynamics, it was found that the β-barrel is stabilised by hydrogen bonds formed by the influx of water in the pore, when the α-helix plug was removed. By measuring the pore size, it was shown that the cargo should be transported through in an unfolded and extended form, addressing questions about the mechanism of autotransporters.

Furthermore, simulations of a truncated BamA structure found that the reduced hydrophobic region between the first and last strands of the β-barrel opened to allow access to the membrane through a lateral opening. This mechanism was observed in a 1 μs simulation, using a truncated *E. coli* ΔP12345 model, with the β-barrel domain embedded in a model membrane.

6.2.1 Theoretical considerations

The 'black box' analogy which is sometimes associated with molecular dynamics can lead experiment biologists to question the validity of these methods [143]. For this reason, it is important to understand at least the basic principles of MD simulations to determine their value for a given study. MD programmes are in fact built upon a solid foundation of physics and are becoming increasingly better at replicating real-world systems.

In theory, quantum-mechanical calculations can accurately describe any physical system, however they are still prohibitively expensive to compute for larger systems. The Hamiltonian operator is used in their place to represent wave-mechanics to a satisfactory level i.e. one at which can reasonably predict the behaviour of a chemical system. Hamiltonian mechanics was developed as a reformulation of classical mechanics that uses a different mathematical method corresponding to the total energy of a system. While a full explanation of the mathematical principles is beyond the scope of this work, a description is provided for reference and to understand the concept of molecular dynamics simulations.

6.2.2 Defining a System

Firstly, consider a system of *N* particles which move under the influence of external forces. Their spatial position can be described as a function of time, where $\mathbf{r}_1(t)\ldots\ \mathbf{r}_N(t)$ is used to denote position and $\mathbf{v}_1(t)\ldots\mathbf{v}_N(t)$ is velocity. The force $\mathbf{F}_1\ldots\ \mathbf{F}_N$ on each particle can be determined by classical mechanics using Newton's second law of motion [144]

$$\boldsymbol{F_i} = m_i\ddot{r}_i \tag{6.1}$$

where $m_{i\ldots}m_i$ are masses of each particle. Newton's equations determine a classical state of system at time *t* with a set of all positions and velocities by assigning initial conditions through solving equation 6.1. However, an analytical solution is not possible in most cases. MD instead uses an iterative numerical approach to obtain an approximate solution using an implementation of the Verlet algorithm [145]. The accuracy of this method is dependent on the time step, Δt, used in the calculations.

The Hamiltonian operator, which separates the total force into kinetic and potential energy functions, is used to describe the energy in the system

$$\hat{H} = \hat{T} + \hat{U} = -\frac{h^2}{2m}\nabla^2 + U(r,t) \tag{6.2}$$

where H is the Hamiltonian operator (total energy) is kinetic energy, V is potential energy, h is Planck's constant and ∇ is the gradient operator. The equation of motion can be derived using the Hamiltonian equation, and by taking the time derivative of each side, determine the classical state by specifying a complete set of positions and their momentums.

6.2.3 Force Fields and Energy Functions

To compute the acceleration and potential energy, the forces on each atom must be assigned. Forces in the system are first separated into two categories: those ***bonded*** interactions i.e. ones which occur between covalently linked atoms within the same molecule, and those ***non-bonded*** which occur between non-covalently interacting atoms in two distinct molecules or between neighbouring atoms within the same molecule.

The molecular mechanics force fields used by CHARMM, GROMACS, and other MD programmes are simplified, parametrised representations of real-world physics. Generally, the energy terms are divided into two categories: bonded and non-bonded interactions (equation 6.3). The first set of terms describes interactions between chemically bonded nearest neighbours, while the second set of terms is used to describe interactions between all other non-bonded atom pairs within a given cut-off distance:

$$U_{CHARMM} = U_{bonded} + U_{non-bonded} \tag{6.3}$$

6.2.4 Bonded Terms

Bonded interactions are measured from bond length (stretching), angle (bending), dihedral and improper dihedral terms.

$$U_{bonded} = U_{bond} + U_{angle} + U_{UB} + U_{dihedral} + U_{improper} + U_{CMAP} \tag{6.4}$$

Bond angle and Urey-Bradley terms (1, 3-pairs) are harmonic potentials each with an associated force constant *K*. The dihedral term represents potential energy associated with rotational motions, or twisting, around a bond. Improper dihedrals are artificial terms which restrict the conformation of tetrahedral molecular geometry in groups with a central atom bonded to three others [146]. The CMAP term is a more recent, protein specific term used to correct backbone dihedral angles (φ) and (ψ) and used to improve conformational properties on the polypeptide structure. Each individual term is as follows:

$$U_{bonds} = \sum_{bonds} K_b(b - b_0)^2 \tag{6.5}$$

$$U_{angles} = \sum_{angles} K_\theta(\theta - \theta_0)^2 \tag{6.6}$$

$$U_{Urey-Bradley} = \sum_{Urey-Bradley} K_{UB}(S - S_0)^2 \tag{6.7}$$

$$U_{dihedral} = \sum_{dihedral} K_\chi(1 + \cos(n\chi - \delta)) \tag{6.8}$$

$$U_{impropers} = \sum_{impropers} K_\varphi(\varphi - \varphi_0)^2 \tag{6.9}$$

$$U_{CMAP} = \sum_{residues} V_{CMAP}(\varphi, \psi) \tag{6.10}$$

6.2.5 Non-Bonded Terms

The non-bonded components of the energy potential function are represented by two terms, the so-called Lennard-Jones, or 6-12 potential and a Coulombic electrostatic interaction term

$$U_{non-bonded} = U_{LJ} + U_{elec} \tag{6.11}$$

$$U_{LJ} = \sum \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r_{ij}} \tag{6.12}$$

By combing all the terms, the CHARMM equation describing potential energy is given as follows

$$U_{total} = \sum_{bonds} K_b(b - b_0)^2 \quad + \sum_{angles} K_\theta(\theta - \theta_0)^2 \tag{6.13}$$
$$+ \sum_{Urey-Bradley} K_{UB}(S - S_0)^2$$
$$+ \sum_{dihedrals} K_\chi(1 + \cos(n\chi - \delta))$$
$$+ \sum_{impropers} K_\varphi(\varphi - \varphi_0)^2 \qquad + \sum_{re\ sidues} V_{CMAP}(\phi, \psi)$$
$$+ \quad \Sigma \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r_{ij}}$$
$$+ \qquad \Sigma[\varepsilon_{ij}(\frac{R_{min.ij}}{r_{ij}}) - (\frac{R_{min.ij}}{r_{ij}})]$$

While other specific decompositions of the terms can be used, a classical additive force-field assumes the general form described above. Many versions of individual force-fields also exist. A set of parameters is used for different atom types, bonds lengths, masses and angles etc. These parameters are usually based on empirical data and are therefore occasionally adjusted based on new experimental findings.

6.3 Computational methods

6.3.1 System setup and the CHARMM-GUI

Simulation input files were generated using the web-based CHARMM-GUI (https://charmm-gui.org) for the full length BamA models of *C. jejuni* and *N. gonorrhoea* in membranes, or via the command line in a Unix BASH shell using the GROMACS 5.1 software package tools for BamD simulations. The *C. jejuni* sequence was modelled in I-TASSER as previously described. The *N. gonorrhoea* sequence was also remodelled using MODELLER using its own determined crystallographic structure (PDB entry: 4K3B). This was done to assign missing atoms not found in the original PDB file. As standard, N- and C- termini were capped with acetyl and methylamine groups to remove charges at the exposed ends of the polypeptide chain. The principal axis of each BamA structure was aligned to the Z-axis and translated 23Å to position the β-barrel at the centre of the membrane. A 10 by 10nm square membrane patch of 3:1 DOPE (1,2-dioleoyl-sn-glycero-

3-phosphoethanolamine) and DOPC (1,2-dioleoyl-sn-glycero-3-phosphocholine) composition comprising of 132 lipid molecules in the top later and 136 in the bottom layer was used. The system was solvated with TIP3 water and sodium chloride ions were placed to a final concentration of 150mM. The CHARMM36 force field was used in all-atom simulations of BamA while simulations of *Cj*BamD model proteins were performed using the GROMOS96-53a force-field, which is optimised for α-helical structures.

The martini22 and martini22p coarse-grained force-fields are also tested. Briefly, this involves converting the PDB using the *martinize* script to build a force field, which allocated four residue-specific atoms to one coarse-grained bead [147]. Water molecules and lipids follow a similar procedure. This considerably reduces the number of particles in the simulation box and massively reduces the number of non-bonded calculations required in each time step, therefore speeding up the overall computation. The input files for coarse-grained simulations were generated manual using the GROMACs 5.0.1 software package.

6.3.2 Trajectory file analysis

Considering the additional need for membrane lipid components, solvent molecules and ions in the system, all-atom simulations of membrane proteins are difficult to handle. Thousands or more snapshots of conformations known as frames, are printed, and inspected manually to observe the behaviour of the system. This was performed using the video mode in PyMOL. Resolving coordinated dynamics related to function though requires both quantitative and qualitative analyses of the data generated. Two popular methods for processing trajectory data are normal mode analysis (NMA) and principal component analysis (PCA). These are both used to reveal major conformational transition occurring within the structures. The Bio3D package features tools used to process trajectory files and perform PCA and NMA.

Normal mode analysis (NMA) is powerful tool for investigating large-scale motions in biomolecular structures. A normal mode is a pattern of motion within an oscillating system in which components within the system move with the same frequency and with a fixed phase relationship. The free motion described by the normal modes takes place at these fixed frequencies. In terms of protein molecules this is understood by regarding the structure not in a single fixed position, but rather as an ensemble of many positions which occur over fixed durations of time. What is often referred to as the "native" state of the protein should instead be represented as an

ensemble of microstates [62]. These states retain the overall fold of the protein which maintain the secondary structure components. However, atomic coordinate positions can fluctuate with fine grained detail and these are often critical to the proteins function.

NMA can be applied to predict functionally related motions i.e. conformational switches related to protein function, often activated, or inhibited by binding of other molecules. NMA always assumes that the largest fluctuations i.e. those with the lowest frequency are functionally relevant [148]. This is because large conformational changes are generally conserved alongside function in the evolution of biological macromolecules. Ultimately, the assumptions on which NMA is based should be derived from experimental techniques and multiple crystal conformers of proteins do in fact confirm that it can be used to predict transitions between states.

NMA is a form of harmonic analysis. It is accurate in some regards as it uses the same force fields designed for MD simulations. However, it assumes that the conformational energy surface can be described as a parabola over a range of temperatures which has been shown to be inaccurate [149] [148]. As such, in physiological conditions, structures do not behave with the simple harmonic motion associated with a single energy minimum as would be predicted in an NMA study. It is also performed in a vacuum, with no consideration for the solvent system. In reality, protein structures cross many barriers with multiple energy minima [150] [151] and as such the NMA approximations, which are usually based on crystal structures with their own limitations, break down when compared to experimental data or results from other computational methods i.e. MD simulations. This is an important limitation to consider when using these tools. Nevertheless, NMA is still useful for investigating general structural properties, it is relatively fast to perform and can be applied in a high-throughput manner to entire ensembles of related structures.

NMA typically takes an atomic co-ordinate set, a force field used to describe the interactions between atoms and a computational programme to perform the mathematical calculations.

The NMA can be broken down into three main steps:

1. Minimization of the surface potential energy as a function of coordinates
2. Calculation of the Hessian matrix; the matrix of second derivatives of the potential energy with respect atomic mass and position
3. The diagonalization of the Hessian matrix yielding eigenvalues and eigenvectors

The eigenvalues and eigenvectors represent the normal modes within the structure which can be used to model of as correlated axis of motion.

Finally, harmonic analysis is a division of mathematics interested in the analysis of complex signals representing the superposition of fundamental sinusoidal wave forms. In recent years, it has been applied to many areas of research and is commonly used in digital signal processing, quantum mechanics, acoustic engineering, tidal wave analysis and neuroscience. Harmonic analysis is another method for decomposing complex coordinated dynamics from trajectory files of MD simulations of proteins. The harmonic properties of the structure can therefore be extracted through NMA analysis or from the MD trajectory and aid in deciphering the structure-function paradigm of biological macromolecules.

6.4 Results

6.4.1 Analysis of all-atom simulations of BamA

The Bio3D R package was used in R-studio to perform RMSD, RMSF per residue, NMA, PCA, and residue cross-correlation analysis (RCCN). Before analysis, trajectory files were converted from the *trr* binary trajectory file to the *pdb* file format before processing. The full-length *Cj*BamA model generated by I-TASSER was simulated in GROMACS 5.0.1 in three replicate 10 ns runs. Optimal performance was achieved on 32 Intel Xeon CPU cores with GPU acceleration using 2 NVIDIA Tesla K80 cards. The results show that in each case, the structures are highly dynamic with an increasing RMSD measurement and do not reach a stable position in the first 10ns of production. Upon closer visual inspection of the trajectory it is apparent that the two N-terminal POTRA domains undergo a greater range of motion and are more dynamic. Analysis of the root mean square fluctuation (RMSF) measurements shows this as the fluctuations are up to 8 times greater in P1 sub-domain region (Figure 6.2).

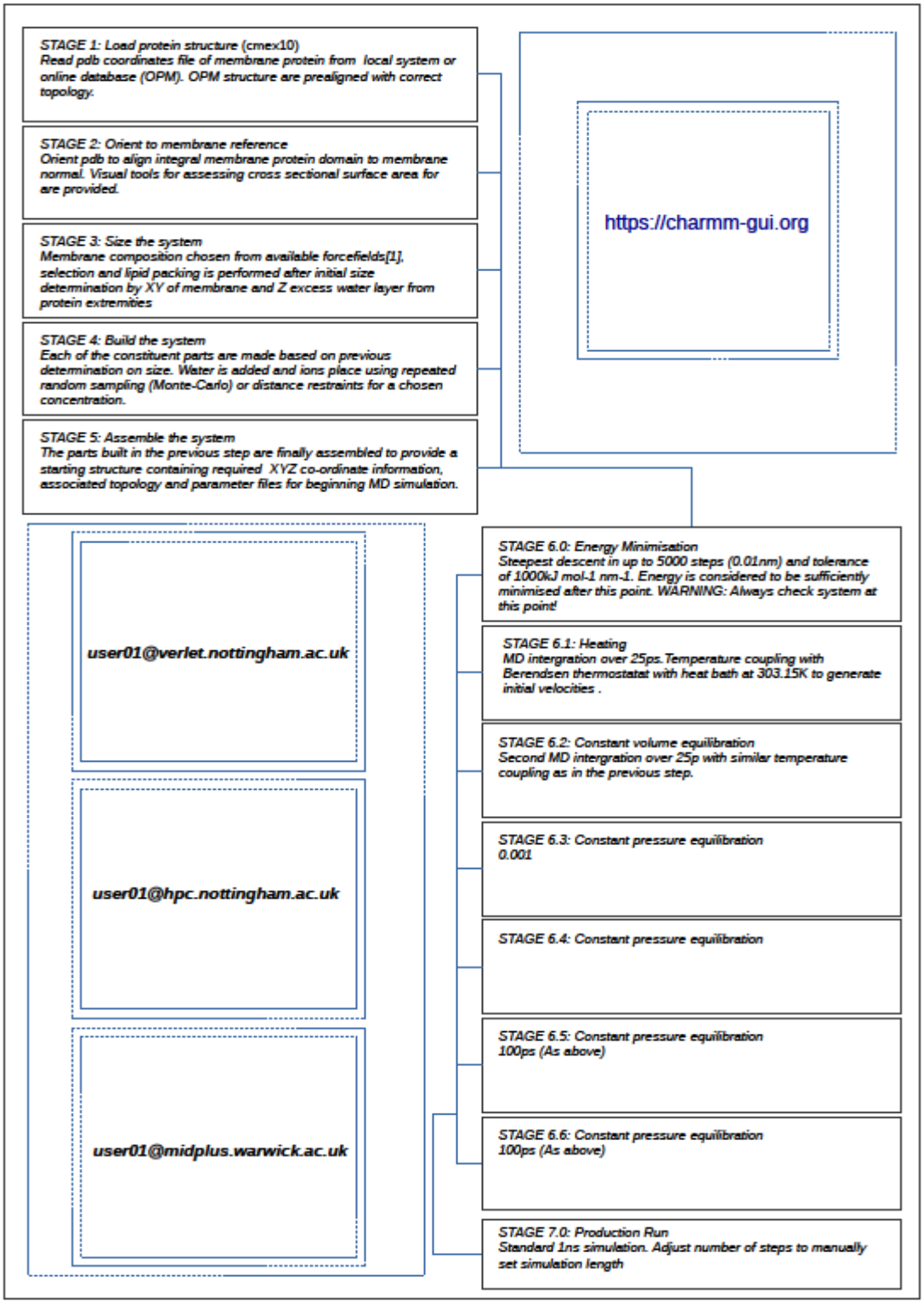


Figure 6.1 Flowchart of MD simulation stages, implemented with the CHARMM-GUI and GROMACs. This procedure was used to run simulations of BamA protein with all-atom and coarse-grained force-fields on high performance computers with GPU acceleration.

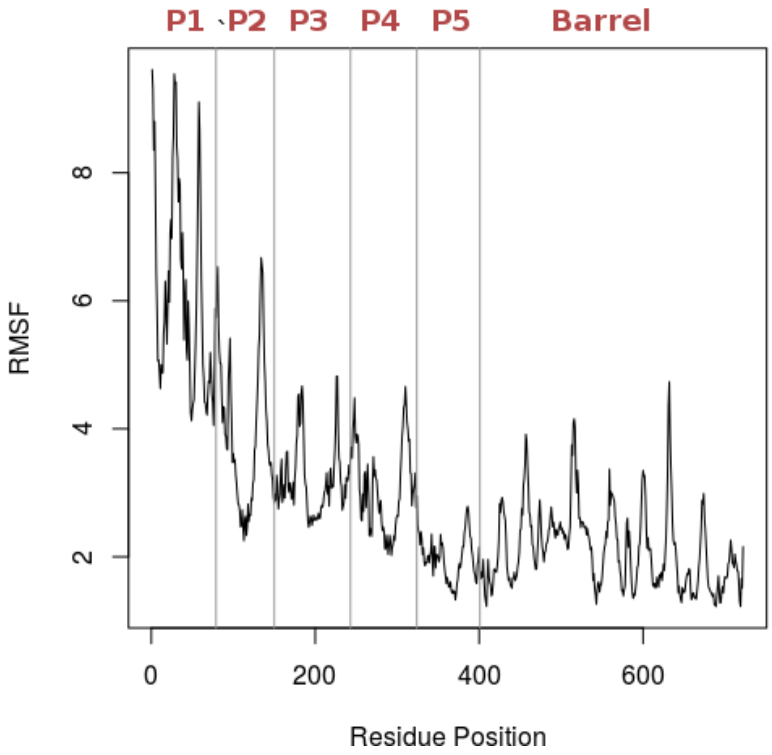


Figure 6.2 Root mean square fluctuation of *Cj*BamA structure over the course of a 10 ns all-atom simulation. The N-terminal POTRA sub-domains P1 and P2 display a higher RMSF associated with enhanced dynamics in this region. A number of peaks are also observed in the barrel domain from residue 400 onwards, occurring at regular intervals along the structure and corresponding to extra- and intracellular loops.

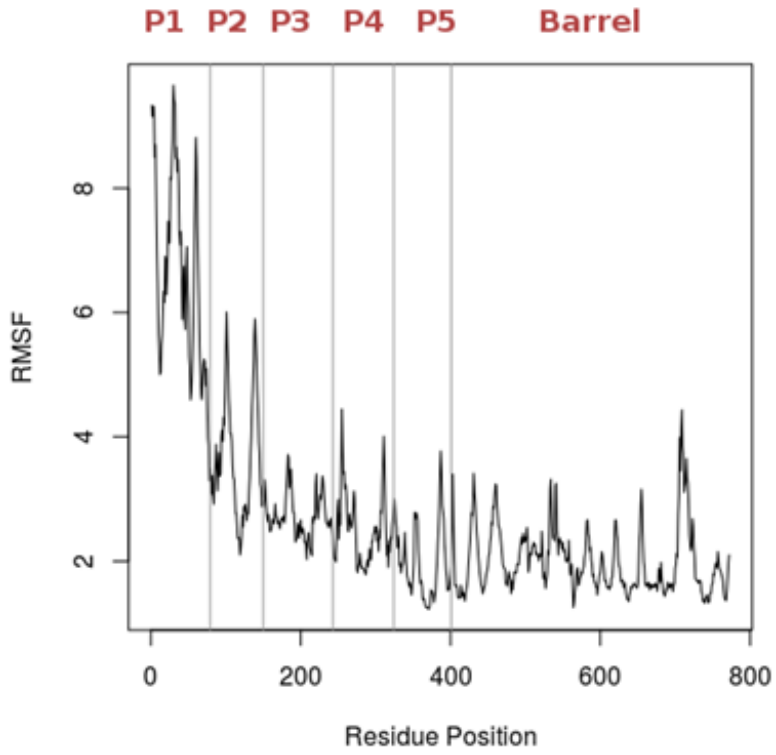


Figure 6.3 Root mean square fluctuation of *Ng*BamA structure over the course of a 10 ns all-atom simulation. Similar to the behaviour of the *Cj*BamA simulation, N-terminal POTRA units P1 and P2 display a high mobility while the barrel appears to be less dynamic overall. Several peaks align to those of the *Cj*BamA model, suggesting these regions may have conserved activity related to function.

RMSF measurements indicating the overall fluctuation in each residue α-carbon are used to identify regions of high mobility in the structure. Interestingly, there is an average linear decline observed in the RMSF from the N-terminal region of the POTRA to the start of the β-barrel around residue, at around position 400. It is apparent that the β-barrel is more stable though, with regular peaks observed corresponding to motion occurring at intracellular and extracellular loop residues between each β-strand.

To determine if this behaviour was conserved in an experimentally solved structure, the atomic co-ordinates of *N. gonorrhoeae* BamA (PDB: 4k3b) was used to perform replicate simulations. The PDB file was first used to remodel the complete sequence in I-TASSER in order replace missing atoms not resolved from the crystal diffraction data. Interestingly, RMSF measurements show that same trend as in *Cj*BamA, with high mobility in the N-terminal region of POTRA domain (Figure 6.3). Furthermore, the β-barrel domain is also shown to be less dynamic overall, compared to the POTRA region.

6.4.2 Trajectory analysis of *Cj*BamA and *Ng*BamA simulations

To determine the overall behaviour and stability of each simulation, an inspection of the system volume, temperature, and energy measurements during the simulation was made after each run was performed to ensure no unusual behaviour had occurred. The overall structural stability of the protein is assessed by measuring the root mean square deviation (RMSD) as a function of time. This indicates the magnitude of movement from the initial structure at each time point. To minimise RMSD bias caused by global drifting in the system, the α-carbon backbone atoms from each frame were first fitted to a reference set. The starting structure (frame zero) was selected as the target reference in the production run using an implementation of the Kabsch algorithm in Bio3D [152]. This procedure was used to minimise the RMSD by translating the coordinate set, computing the covariance matrix, and optimally aligning the rotation matrix to generate an aligned measure of the protein dynamics. Hence, it first aligns the structures to one centroid position in space, removes rotational degrees of freedom and then measures the evolution of the backbone over time. The resulting data is therefore reduced to a centroid, free from tumbling rotation, drifting motions and ready for structural analysis.

To assess the conservation of dynamics in BamA protein simulations and determine confidence in the homology models used, a set of runs were performed using a system reconstructed using

identical build parameters, with *Ng*BamA remodelled in I-TASSER. In both cases, the replicate simulations show varied behaviour (Figure 6.4). Although the initial conditions were preserved in each replicate, each one shows a different behaviour. However, an initially steep upward trend in RMSD is made in each case as the structures move away from their starting structure. The rate of change in two *Cj*BamA runs (Figure 6.4, A and C) are similar and after around 6 ns the rate decreases. Conversely, the second run (B) alters its trajectory sooner, after approximately 2.5 ns when the magnitude of acceleration is approximately equal but opposite, indicating that the structure is moving closer back toward the conformation of the first frame. A similar behaviour is seen in one *Ng*BamA run (E), where the rate change event occurs after approximately 6 ns. In the other two *Ng*BamA runs, varied behaviour is also observed. The first run shows that the structure stabilises within a 1 Å range, after 2.5 ns. After the same time into the final *Ng*BamA simulation (E) the structure transitions to a new phase, rapidly moving away from the initial conformation to reach a final RMSD over 7Å away from the initial structure.

In all cases the rate of change of RMSD does not appear to increase exponentially, perhaps indicating that the protein structures are moving dynamically out from restraints imposed in the crystal template and towards a more natural state - confirming that the structure is not rigid.

6.4.3 Extended P23 loop interactions with membrane

Upon visual inspection playback of the simulations, an extended loop in the *C. jejuni* structure opens to interact with the membrane (Figure 6.5). Examination of the amino acid sequence showed that two hydrophobic tryptophan residues are in this region, and these amino acids are well conserved amongst different organisms. In frame 1 (Figure 6.5, A) the loop starts in a closed position. After 4 ns (B) it has partially opened and by 8 ns (C) it has begun to interact with the membrane. By anchoring in the membrane, this extended disordered region might act to provide stability in the absence of other auxiliary lipoproteins BamBCE which are not present in the *C. jejuni* model of BAM.

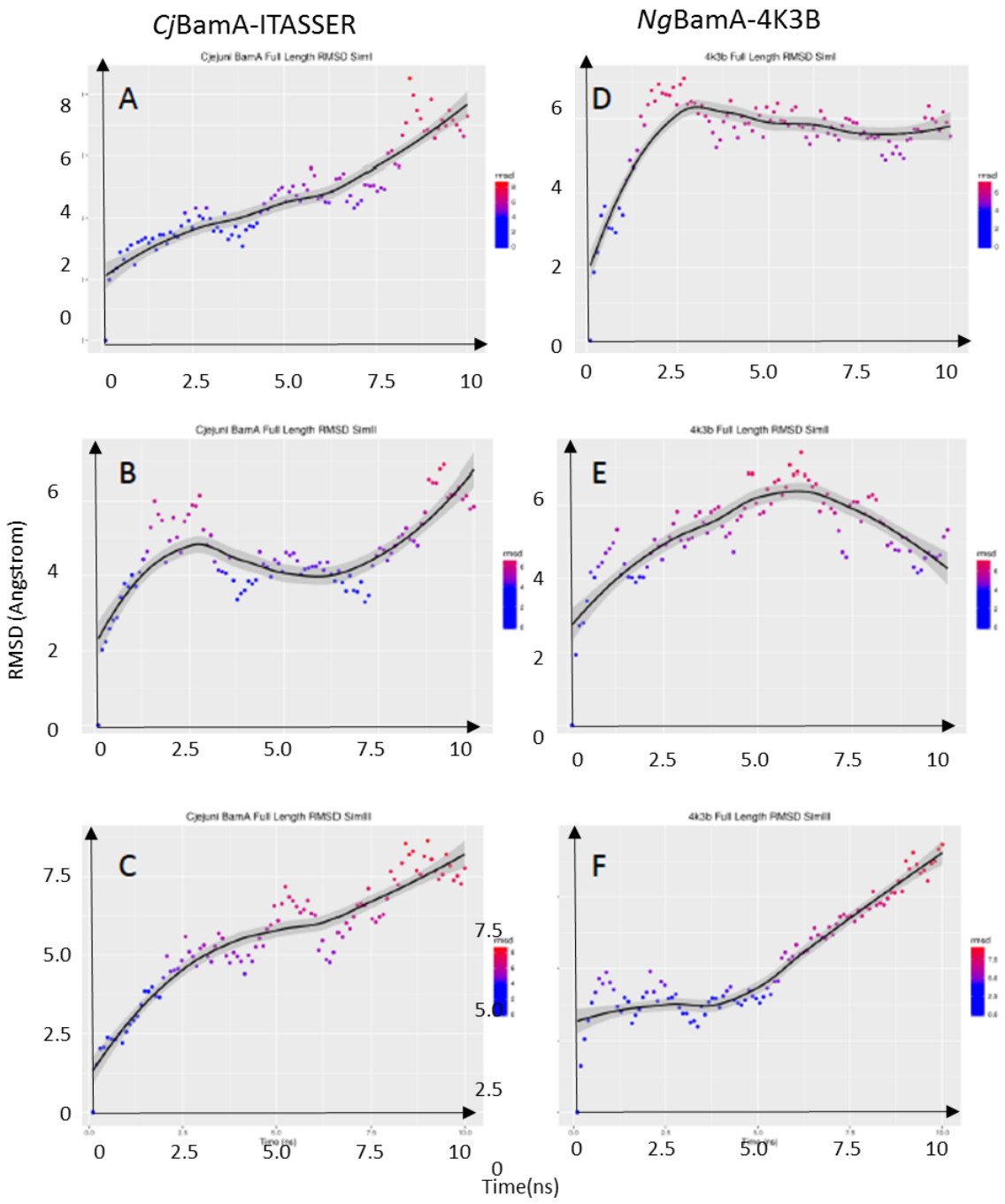


Figure 6.4 Three replicate 10ns simulations of *Cj*BamA I-TASSER model and *Ng*BamA 4K3B show varied behaviour. In the first *Ng*BamA (D) simulation, a stable state is reached after approximately 2 ns. In two (B and E), the trajectory appears to return closer to the starting conformation, while in others the structure evolves away from the original pose.

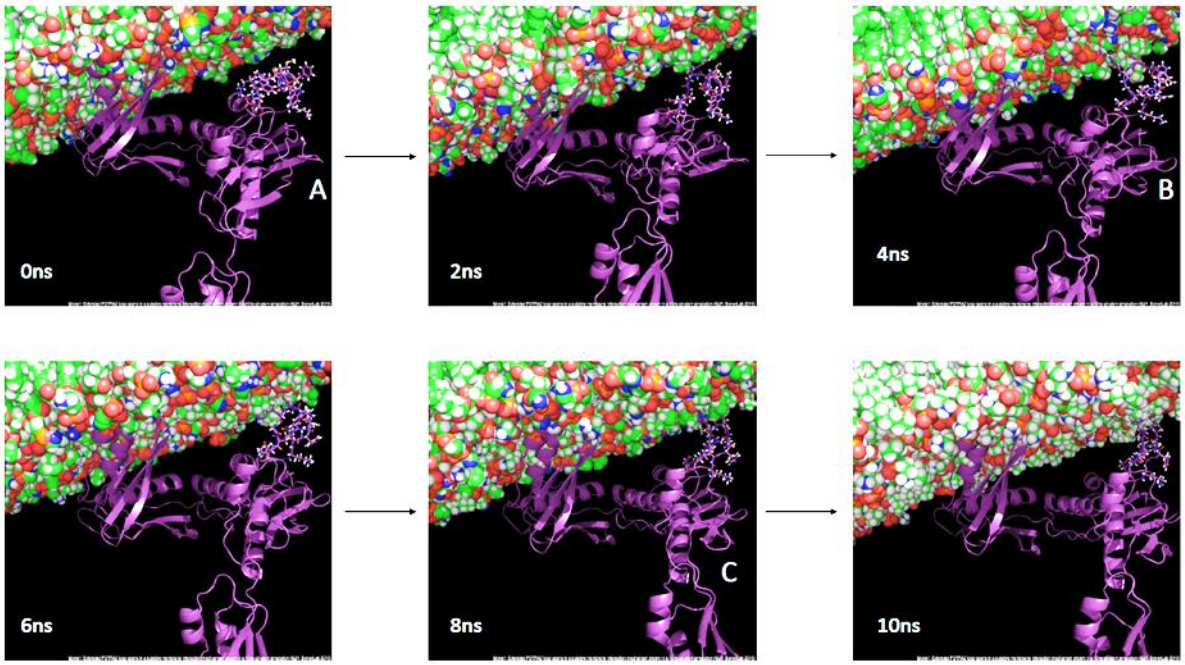


Figure 6.5 BamA POTRA domain loop interacts with the bilayer. After approximately 2 ns (A) the extended loop of P2 begins to open and by 4 ns (B) it has started to make contact with the bilayer. Two aromatic tryptophan residues then rotate to face the membrane and the relevant sides chains start to embed after around 8 ns. Please see the additional animation files (available upon request)

While these loop residues are also found in other organisms with additional lipoproteins, it could indicate a conserved feature which is used to anchor the POTRA domain to the membrane when the auxiliary components are not present, to facilitate OMP assembly in a reduced BamAD system. As seen in the literature BamBCE are non-essential components which supports the idea of an altered mechanism relying on the BamAD components alone. Although this requires further investigation, it could explain how BAM complex can function with the minimal set of essential BamA and BamD components.

6.4.4 Extended simulation of *Cj*BamA

Short 10 ns simulations were used to determine the dynamics of two different BamA models. However, to observe how the *Cj*BamA model behaved over an extended period, the simulation of the I-TASSER model was extended to 111 ns (Figure 6.6). Again, initially the structure rapidly moves away from the starting pose. After a linear increase in RMSD, at approximately 60 ns the rate of change starts to lessen, and large oscillations begin to emerge.

After 90 ns, the overall rate of change in the structure begins to decrease, however it cannot be seen if the model reaches a point of regular periodic motion within the 111 ns time frame. These results suggest that the structure generated by I-TASSER could be representative of a highly dynamic and functional protein. This is inferred from the pattern of oscillations which begin to emerge after 50 ns. During the first half of this simulation, as observed earlier in shorter 10 ns simulations, the protein dynamics are faster and rapidly move away from the starting conformation, possibly due to steric hindrance imposed from the crystal structure template. After an initial adjustment period, the BamA structure could then begin to relax and start regular oscillations with augmented motion at the N-terminal region of the POTRA domain. How this may drive OMP assembly remains unclear from these simulations, but they reveal some important aspects about the model quality - one important observation being that the change in RMSD over time starts to become less. This is also observed by measuring the RMSD during equilibration stages, which further showed that the structural trajectory gradient appears to flatten over the course of the simulation.

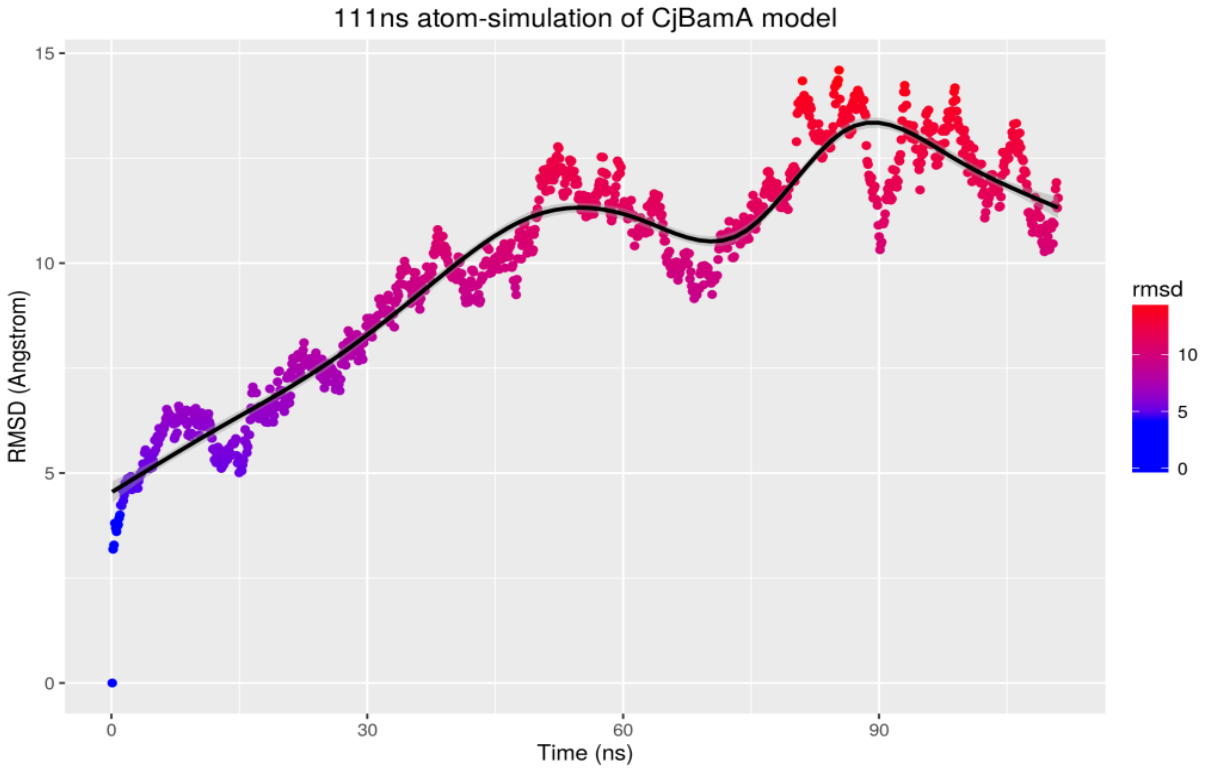


Figure 6.6 Extended 111 ns simulation of I-TASSER *Cj*BamA run in NAMD. Minor oscillations occur with an overall upward trend as the structure moves away from the starting confirmation, predicted using I-TASSER. The average rate of change is initially steady but after 50 ns the trend is shown to reverse, returning to a point less than 10 Å from the starting structure before once again continuing its upward trend at 70 ns. After 90 ns the trend is reversed again. These events support the notion that the protein goes through a cyclic process to assemble ascent OMPs.

6.4.5 Analysis of internal domain dynamics

Although the full structure was shown to be highly dynamic, to assess the dynamics with each local sub-domain, RMSD measurements were made to assess the stability of each region with respect to themselves (Figure 6.7). This was done by splitting the *pdb* file into respective regions and fitting them each back to their own starting reference set before calculating the RMSD, using a script implemented in Bio3D. The overall structure is shown to be dynamic when the RMSD measurement is made but by aligning the domains against themselves it is seen that they maintain relatively stable conformations. This result suggests that the predicted structures of each POTRA domain are stable and could be feasible in a functional model. A greater level of motion is again seen in the N-terminal POTRA P1 domain, as previously observed and the RMSD is seen to cycle over 20 ns periods, which may suggest a cyclic nature in the dynamic behaviour of this protein.

The results of PCA for both *Cj*BamA and *Ng*BamA simulations show that many accessible degrees of freedom can be extracted. Plots of PC1 versus PC2 in both cases indicate the structures sample distinct conformational regions. Although the number of atoms in the system makes it difficult to determine unique states, this method shows that each structure could behave in a similar manner, with essential motion related to function

A high level of rotation degrees of freedom associated within the system make it difficult to concisely describe the motion of the protein in space. Residue cross correlation analysis is used to provide a visual representation of residue pairs which are moving in phase with one another in correlated motion and highlights some of the important global interactions occurring within the structure.

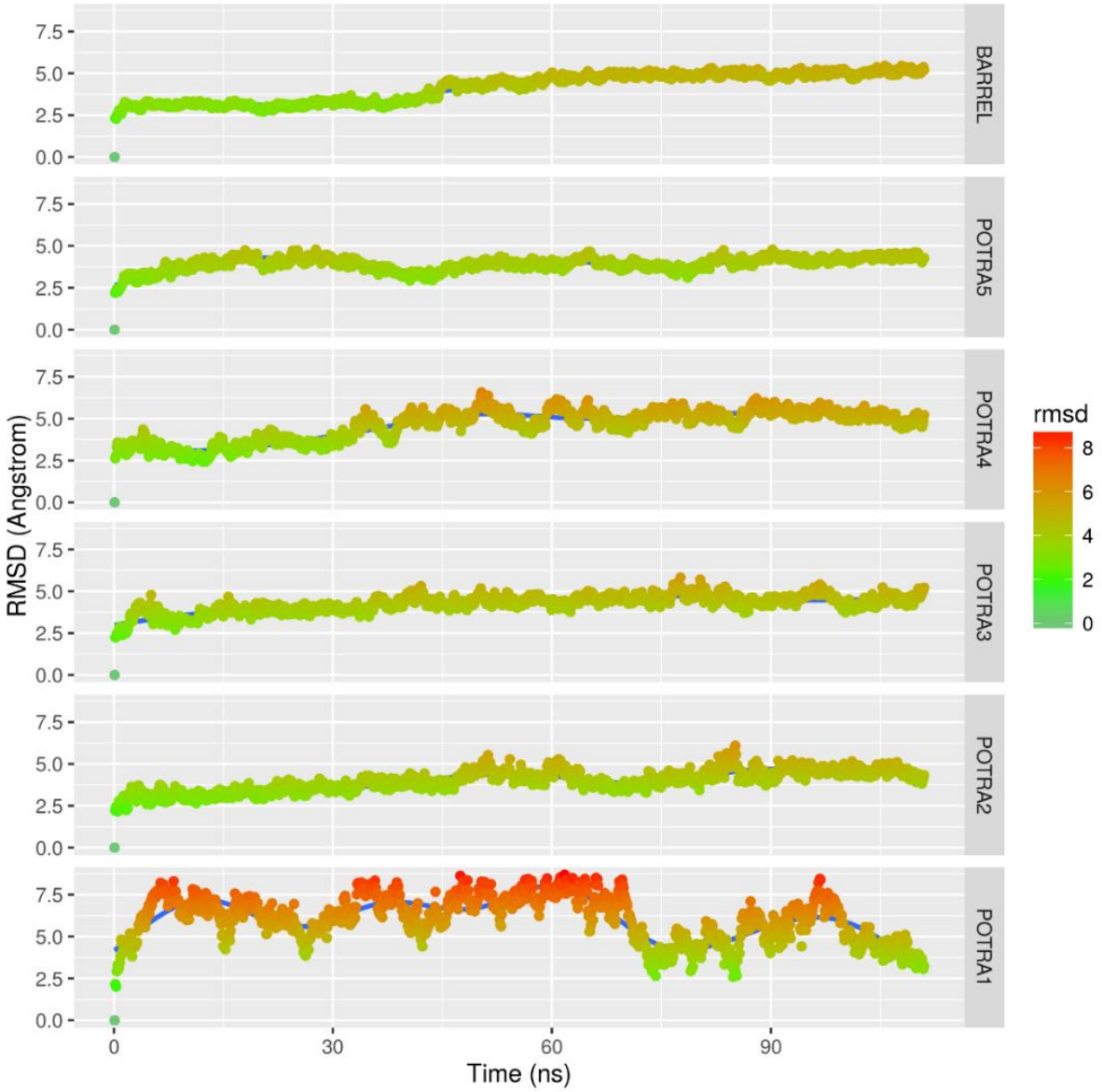


Figure 6.7 Internal dynamics of BamA sub-domain within a 111 ns simulation. This figure shows that the N-terminal POTRA domain, P1 is the most dynamic, while P2345 are much more stable. The C-terminal, P5 subdomain appears to be more stable, suggesting that it could be supported by interaction with the membrane and β-barrel domain. After 40 ns the β-barrel transitions from one stable state to another, approximately 2 Å away, which corresponds to increased activity in P4 and dampening in P1. These results support the enhanced dynamic behaviour of the N-terminal POTRA units described in literature [70]

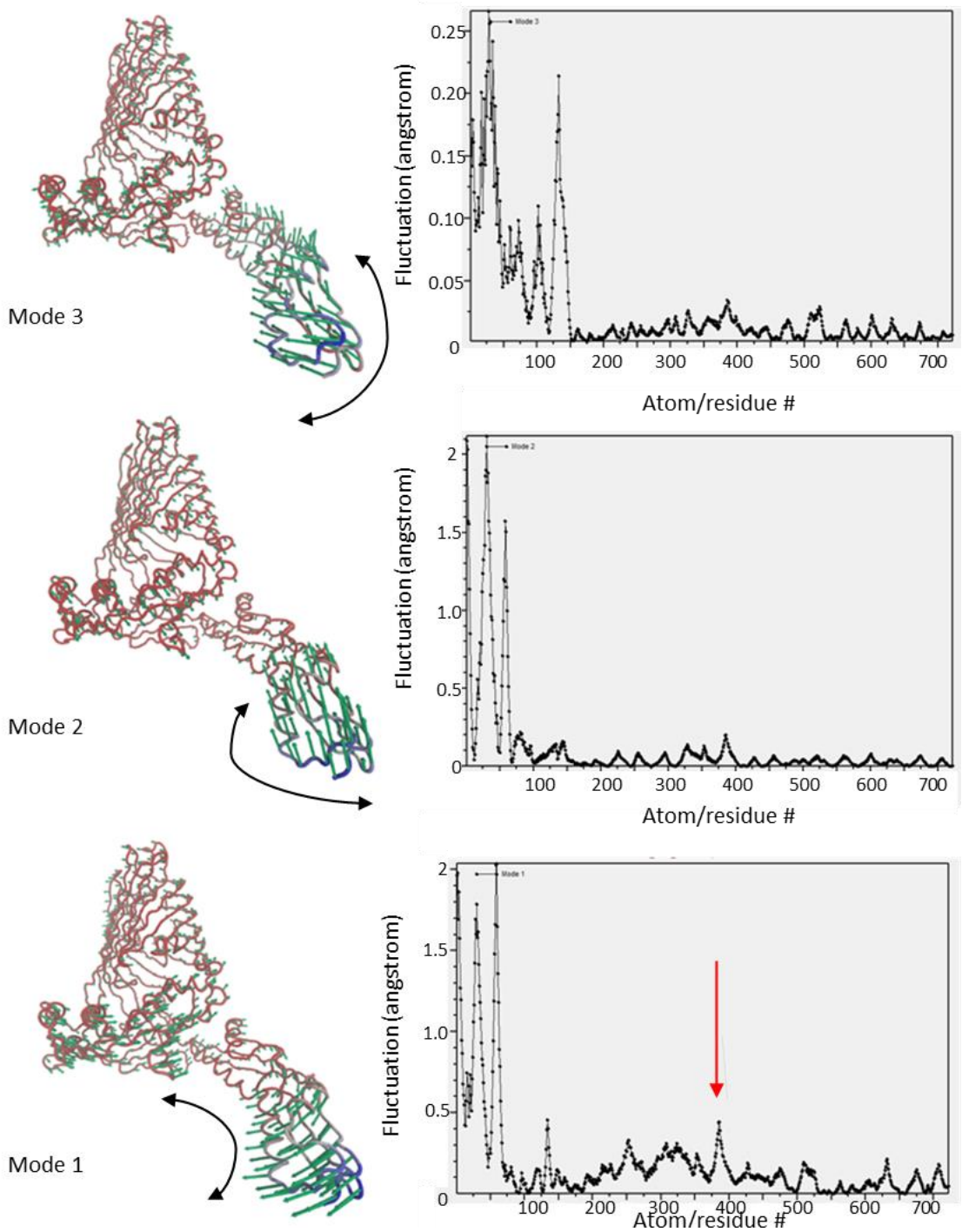


Figure 6.8 Normal mode analysis of *Cj*BamA. High levels of mobility are detected with NMA in the N-terminal region of the POTRA domain, while the β-barrel appears to be relatively stable. Flexibility in the linker region (red arrow) is also shown and enhanced in the first mode calculated.

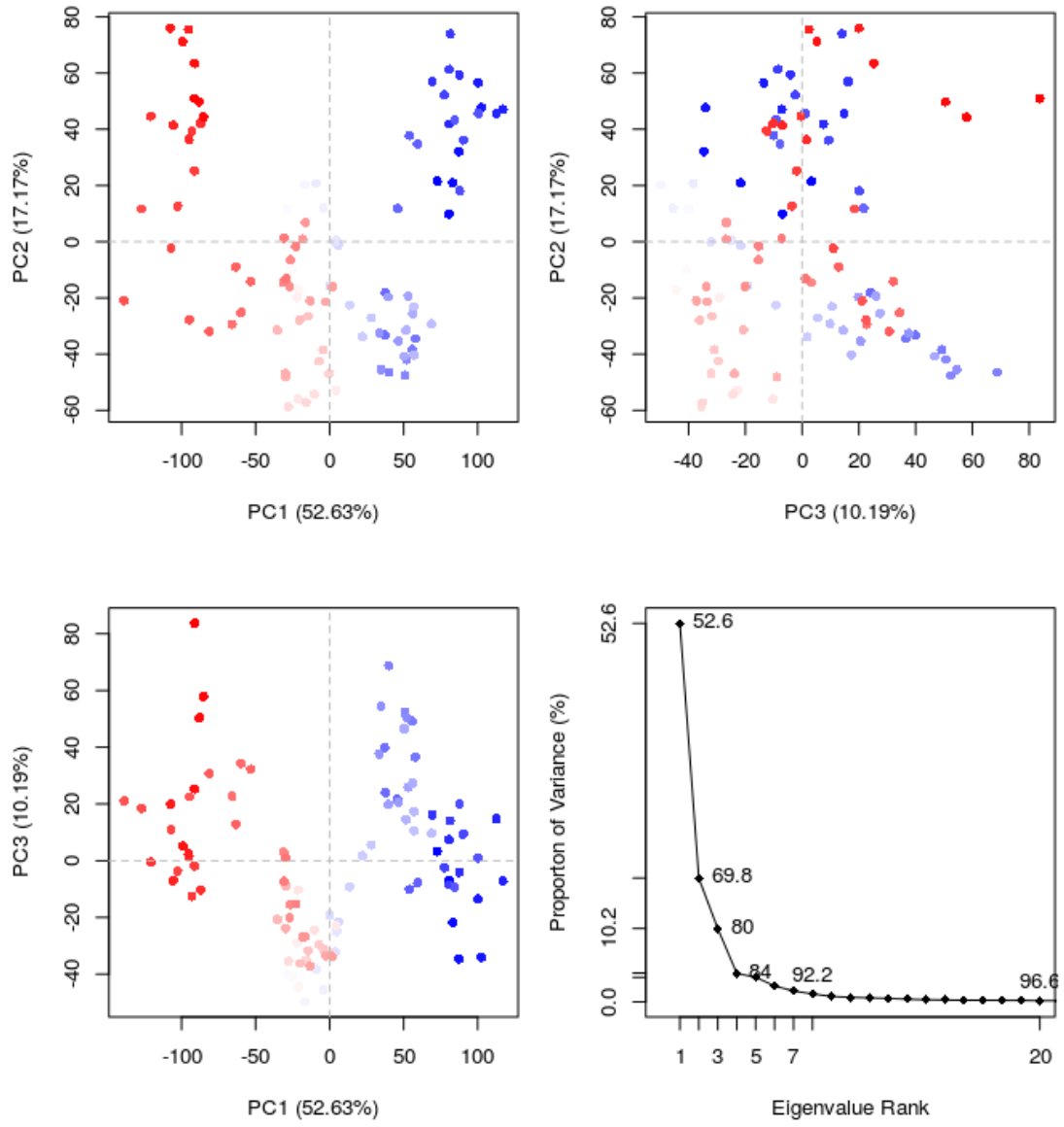


Figure 6.9 Principal component analysis of *Cj*BamA. Three principal components are plotted show conformations sampled by the protein structure during simulation. A high level of symmetry is observed between the PC1/2 and PC1/2 components extracted indicating that a few principal axes of motion are present.

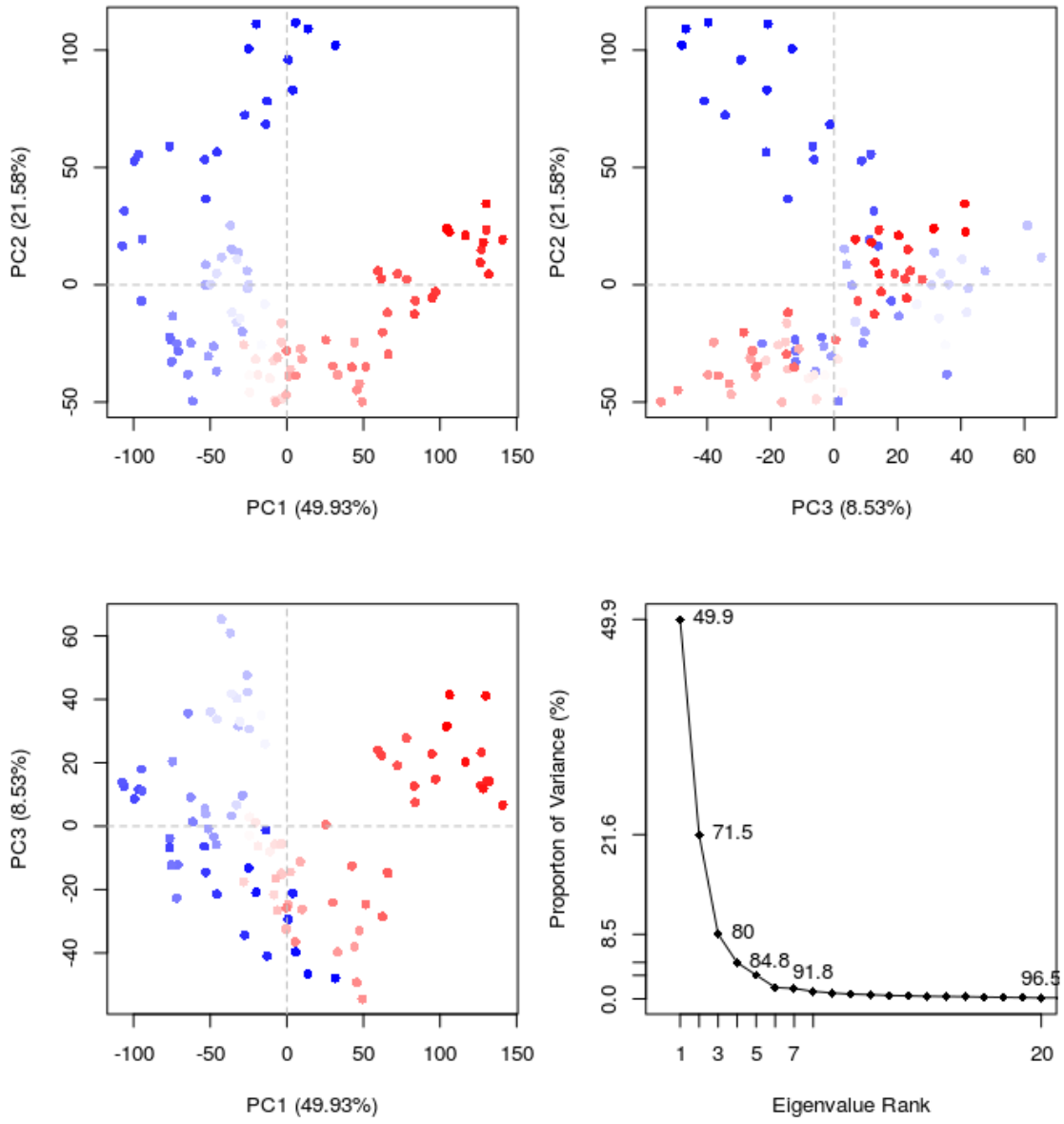


Figure 6.10 Principal component analysis of *Ng*BamA. The structure is shown to sample a number of different spatial regions within the course of a 10 ns simulation. There is also a strong level of symmetry in the clusters, which can be projected by mapping the coordinates of each principal component in PyMOL (session and animation available on request).

A high level of rotation degrees of freedom associated within the system make it difficult to concisely describe the motion of the protein in space. Residue cross correlation analysis is used to provide a visual representation of residue pairs that are moving in phase with one another in correlated motion and highlights some of the important global interactions occurring within the structure.

To better visualise some of the complex dynamics which occur during these simulations, a cross correlation network for *Cj*BamA was generated using the Bio3D package from a 10 ns trajectory (Figure 6.11). As seen previously in the signal map generated from co-evolutionary signals (Chapter 4, Section 4.3), distinct groups of local interactions can be seen for each POTRA subdomain. Long-distance contacts are also observed throughout the structure but are concentrated at the membrane normal. Similarly, an analysis of an *Ng*BamA simulation shows the same features (Figure 6.12). These results indicate that there is a strong dynamic structural correlation which is common in both models – especially between residues 400 and 600 which shows how the lower portion of the β-strands are interacting.

At first, unlike in RMSF and RMSD analyses, the two N-terminal POTRA subdomains appear to be less active near the backbone – which runs diagonally bottom left to top right - across the image. However, upon closer inspection, broad regions of opposing motion are seen in the wide bands of out of phase correlation across the bottom of the figures (the intensity is shown in blue and pink, respectively). More specifically, this indicates the N-terminal POTRA sub-domain unit is moving in a negatively-correlated motion with P345 units, while moving in-sync with regions around the β-barrel – an event which is also seen in the *Ng*BamA correlation analysis.



Figure 6.11 Residue cross correlation network of a *Cj*BamA simulation. A cross correlation map generated using Bio3D, shows regions in green with a positive correlation where residues are moving in phase, while regions in blue are out of phase. Most of the in-phase motions occur close to the backbone, while many long range opposing interactions can be seen.

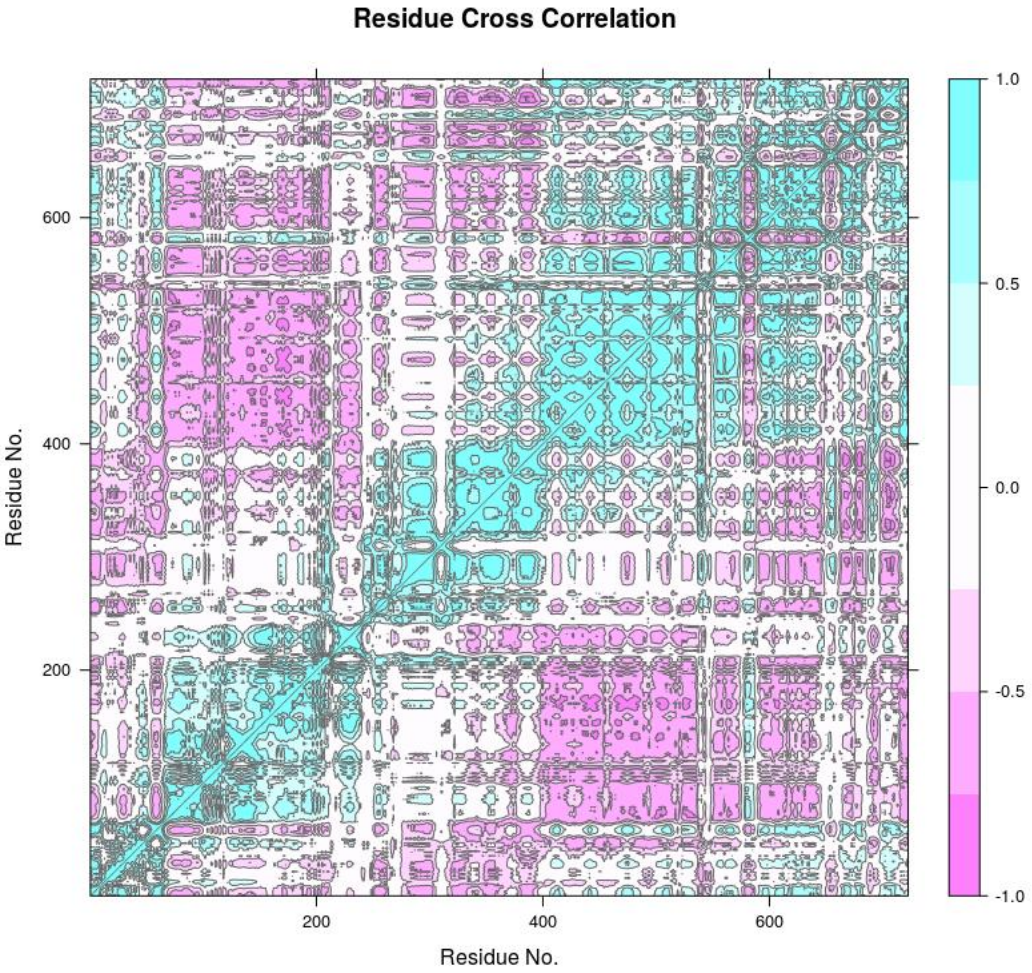


Figure 6.12 Residue cross correlation network of a *Ng*BamA simulation. A cross correlation map generated using Bio3D, shows regions in green with a positive correlation where residues are moving in phase, while regions in blue are out of phase. Most of the in-phase motions occur close to the backbone, while many long range opposing interactions can be seen.

6.4.6 Course-grained simulation of full-length BamA

Coarse-graining methods were used to conduct molecular dynamics simulations of two models of BamA using the Martini-22p force field. The results of coarse-grained simulation over 100 ns show that the structure behaves in a similar manner as seen in all-atom CHARMM simulations. However, the fluctuations are much greater that those previously measured (Figure 6.12). The N-terminal POTRA domain is more dynamic as the coarse-grained force field appears to allow for much greater levels of motion.

Computationally, these simulations are much less intensive though due to the reduced number of particles after coarse-graining and simulations up to 1 μs were feasible. However, further unconstrained motions were observed, and the protein collapsed towards the membrane after a few hundred ns into the simulation (Figure 6.13). This was seen in both the *Cj*BamA model shown, and a coarse-grained 4K3B structure indicating that the force-field used in these simulations requires addition constraints.

6.4.7 Simulation of BamD models

Using the recently discovered sequence of *C. jejuni* BamD, three models were generated using I-TASSER, MODELLER and PHYRE2, as described previously (Chapter 4). To assess the structural integrity and behaviour of these structures, replicate simulations were conducted using GROMACS with the GROMOS-56a force field, optimised for α-helical components (Figure 6.13). The results show that the model generated by MODELLER behave worst i.e. it is the least stable, reaching a position over an RMSD of 10 Å away from the initial pose.

The PHYRE2 and I-TASSER models are more stable, displayed more regular dynamics, and both cycle between several conformational states after initially moving away from the starting position. After approximately 3 ns, the I-TASSER model reaches a peak RMSD value of 6 Å, while a similar state is reached after 6 ns in the PHYRE2 simulation. After this period, the I-TASSER model then reaches a more stable state in which the structure starts short periodic oscillations. However, the structural trajectory of BamD generated in MODELLER does not behave in the same manner, indicating steric hinderance.

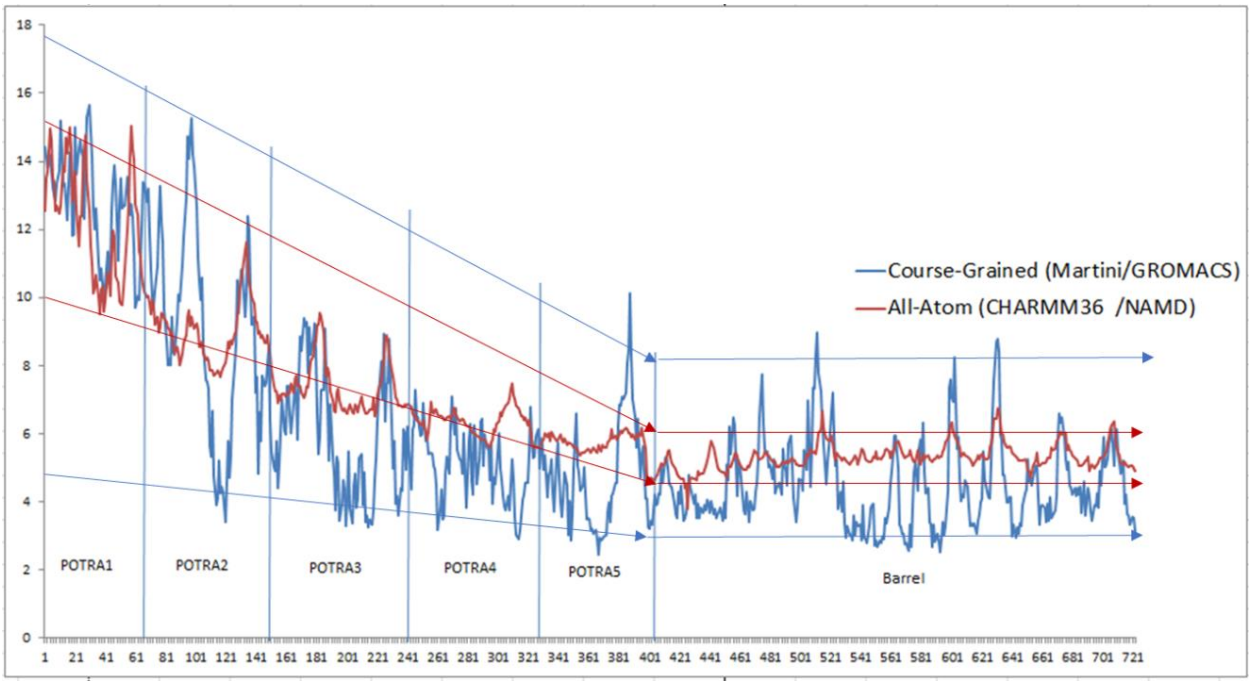


Figure 6.13 Root mean square fluctuation of all-atom and coarse-grained BamA. Over 100 ns, high mobility is observed in both all-atom and coarse-grained models at the N-terminal POTRA domain. The amplitude of fluctuations observed is much greater in the coarse-grained simulation though, while the atomistic force-field behaves better, is less dynamic overall. A number of peaks are aligned in the POTRA domain and β-barrel, corresponding to loop regions which exhibit enhanced mobility as seen previously (Figure 6.2).

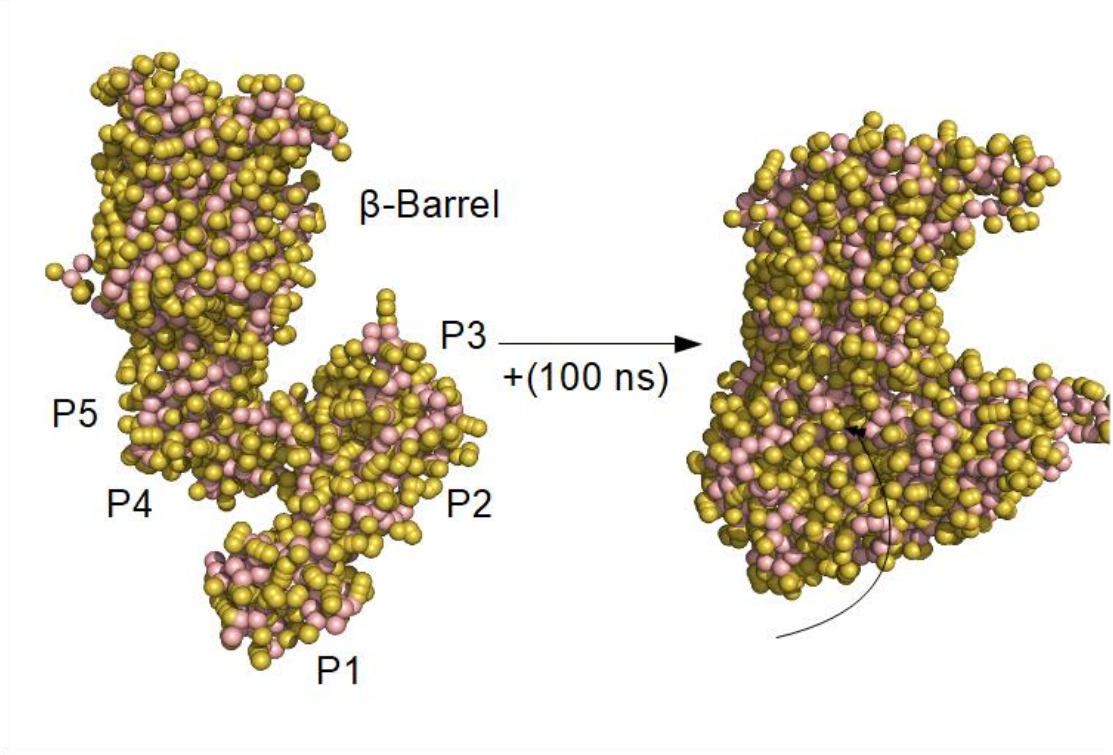


Figure 6.14 GROMACs simulations are used to critically assess the quality of the generated model from MODELLER, I-TASSER and PHYRE2. The BamD model generated with MODELLER simulation indicates poor model quality, while the trajectory analysis of models from I-TASSER and PHYRE2 both show that the theoretical structures are more stable.

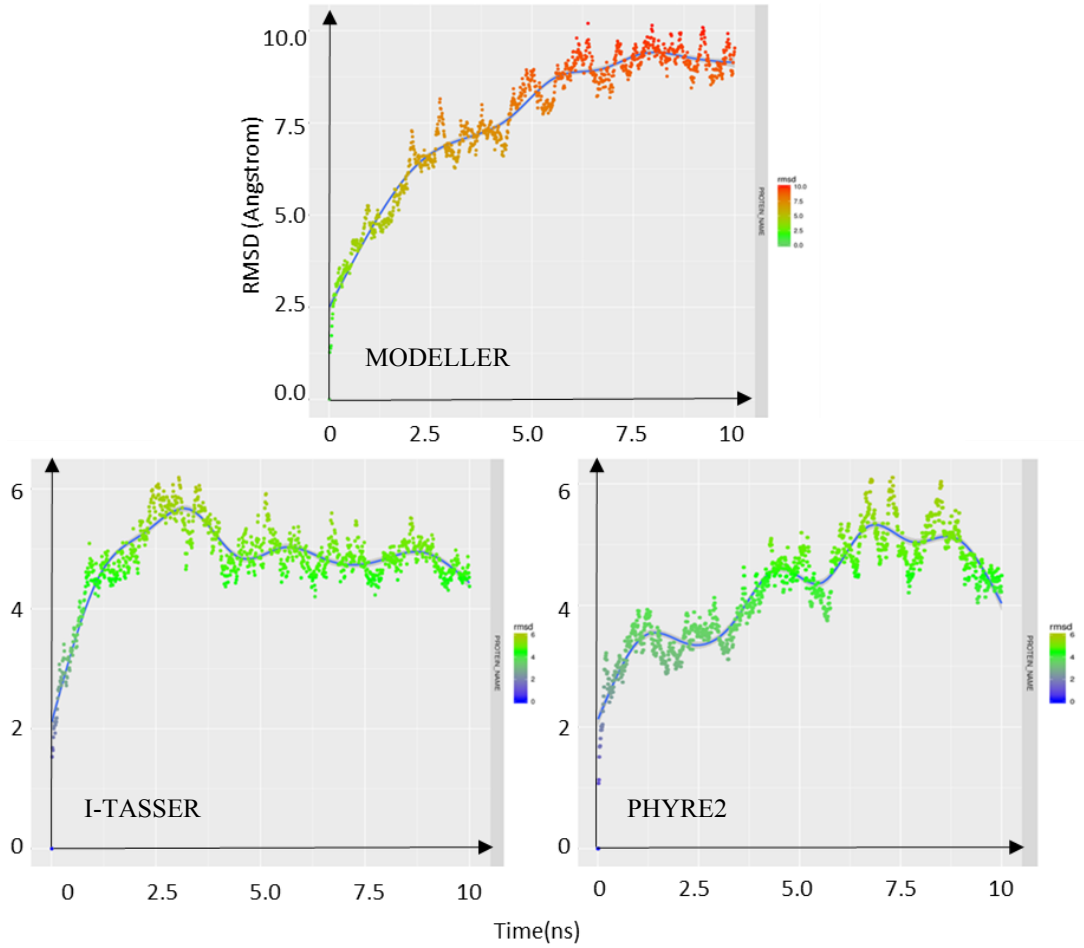


Figure 6.15 GROMACs simulations are used to critically assess the quality of the generated model from MODELLER, I-TASSER and PHYRE2. The BamD model generated with MODELLER simulation indicates poor model quality, while the trajectory analysis of models from I-TASSER and PHYRE2 both show that the theoretical structures are more stable.

6.5 Discussion

The structure and mechanism of β-Barrel assembly in Gram-negative bacteria has been studied previously using a range of experimental techniques. However, molecular dynamics programmes have been increasing popular and proven to be a useful platform for providing computational approaches to study protein dynamics and behaviour in systems like this. Although complex protein assembly pathways are still difficult to study, these methods have a potential to investigate complex mechanics that might otherwise be unfeasible to observe practically, using traditional approaches. Furthermore, recent advancements in hardware have made it possible to run simulations over more meaningful time-scales and for larger systems.

Preliminary 10 ns simulations conducted in a comparative study between *Cj*BamA and *Ng*BamA showed the dynamic behaviour of these proteins appears to be conserved in some manner. Although this is considered a relatively short time period when compared to simulation time scales achieved using the state-of-the-art hardware technology, it was sufficient to observe shared dynamics and in one case, for the remodelled *Ng*BamA-4k3b, the structure looked to reach a stable position within this time-frame. Although MD results should be viewed with uncertainty when lacking experimental evidence to support, they have been shown to be suitable for replicating observations made previously about the dynamic nature of BamA. The results shown here are in accordance with previous observations made about the dynamic behaviour of BamA through solid state NMR, specifically, enhanced motion and flexibility in the POTRA domain [153].

Previous reports of molecular dynamics studies of a BamA β-barrel domain also highlighted instability at the reduced hydrophobic region between strands β1 and β16, which is thought to be an integral part of the mechanism of insertion. This was seen in a truncated model of an *E. coli* BamA β-barrel domain; however no such observation could be made in the simulations here. When this work was undertaken, an all-atom simulation of the full length BamA structure had yet to be reported. However, results have been recently published for an MD study using a hybrid structure, which was carried out for the full-length *E. coli* BamA protein [154].

Both previously published MD studies reported the use of an *E. coli* homology model, but contradictory findings were made about the dynamic opening of the β-barrel domains lateral opening mechanism. Critical assessments of the models generated were not available and, as such, it was not clear how this impacted the results of the simulations. These details were particularly important for the second study as a hybrid structure, modelled on multiple templates and manually

recombined, was used. Within the first 100 ns of the simulation of this hybrid model, the POTRA sub-domains had dramatically repositioned up towards the membrane. However, this was not seen in simulation of *Cj*BamA simulation after 100 ns. Poor stitching between the components could lead to excessive force around a hinge region causing unusual stress in the structure which subsequently moves away from its original position before returning to a better energetically favourable position in a theoretical unfolding-refolding event. Concerns regarding the influence of the neighbouring unit cells and interaction crossing the periodic boundary conditions were also not fully addressed.

The time scale for these events, based on those simulations seen, would be on the order of microseconds, assuming the structure was not caught flattening on the membrane for a prolonged period. This might be further investigated through theoretical measurement of the rotational forces on important hinge residues from the model and correlated to those found in crystal structures of full length and truncated BamA proteins. While it was recognised that the crystal structure templates used may not be representative due to the conformation adopted, the simulation data should be fitted to the experimental evidence at hand, rather than dismissing crystallographic findings.

Furthermore, it is shown in supplementary material of this article that the N-terminal tail of the POTRA domain interacts with the periplasmic-facing underside of POTRA4 in two of the three replicas performed. This effect could easily have adversely influenced the outcome of the results when considering the motion of a single monomeric unit. The solvent thickness of the unit cell would ideally be increased curtail this effect and to see if the observed POTRA-membrane interaction event is recreated. Upon inspection of the BAM complex crystal structure, the location of lipoprotein accessory units appears to provide structural support to the BamA core, in particular the pinning provided by from BamB through BamA and BamD. Although the complete opening and flattening of the POTRA region along the membrane observed in the simulations described showed inherent flexibility between POTRA domains, it radically moves away from the current proposed mechanism and raises many more uncertainties. With such a gross rearrangement, it would take considerable energy to reverse the cycle.

On the other hand, both short and long simulations of BamA in a membrane presented here show that while the POTRA domain is highly mobile, the motion is co-ordinated while being constrained within a relatively small space. The interaction of the extended P3 loop and embedding

of aromatic residues in the membrane as a putative mechanism for stabilisation is also considered to oppose a theory of large global rearrangements in the POTRA regions.

Increasing simulation length and expanding the study to new, larger systems involving more components will provide more detail on the mechanism. Currently, the growth of MD is benefitting from faster computers and more accurate force-fields. Advances in hardware such as the MDGRAPE-3 machine at RIKON, Japan, which outperformed the then TOP500 leading IBM Blue Gene/L system have also made MD simulation capable of achieving much longer time scales for larger systems. Since then, other system such as ANTON, developed by the DE Shaw Group (USA), have surpassed previous milestones in performance by using custom designed application specific integrated circuit (ASIC) chips which are specialised for MD applications [155]. MDGRAPE-4 has also now in development, seeing in a new era of molecular dynamics and protein folding simulations [156].

In summary, the mechanism of assembly appears to involve co-ordinated motion around several axes in the BamA structure, which have not previously been shown. These motions are occurring predominantly in the POTRA domain, specifically the N-terminal P1 subdomain around the P12 linker, which has enhanced flexibility. Although it has not yet been feasible to simulate the system with other essential components, namely BamD and any chaperones, and a nascent OMP cargo, the analysis of MD trajectories has proven to be a useful tool for studying the behaviour and structural integrity of both experimentally solved and model BamA proteins.

Chapter 7. Conclusions and Future Work

In this work, we sought to understand the structure, function, and mechanism of β-barrel outer membrane assembly in the Gram-negative pathogen, *C. jejuni*. The sequence of *C. jejuni* BamA has been used for recombinant expression, secondary structural analysis using circular dichroism and in computational methods used to understand structure-function-mechanism relation in this essential pathway. Synchrotron radiation circular dichroism was used to gather secondary structural information of purified *C. jejuni* BamA and confirmed the structural prediction made from template-based homology models of BamA proteins. A dominant β-sheet content associated with the transmembrane β-barrel domain is shown, in agreement with known structure of BamA protein structures solved at high resolution using XRCD. Thermal protein denaturation was also used to determine the melting transition temperature at which the structure unfolded in two unique detergent buffer conditions. It was shown that the Zwitterionic SB12 detergent solution was able to maintain folding up to 20°C higher than in the non-ionic DDM solution. Using SDS-PAGE, the heat modifiability of the protein in these two solutions was confirmed, showing that the SB12 micelle protects the protein from thermal denaturation above 50°C.

7.1.1 Optimisation of protein production

In summary, initial over-expression was found to be successful for obtaining high yields of relatively pure inclusion bodies from small liquid cultures. However, the cumbersome task of refolding was a time-consuming process involving the use of strong chemical denaturants as well as being inherently difficult to monitor. Furthermore, purification by IMAC with FPLC was not achieved after attempted refolding indicating the hexa-histidine tag was either masked or buried in the barrel. Due to its proximity to the barrel, the C-terminal may have even completely destabilised the fold at this region. Alternatively, targeted expression and isolation from the host cell membrane fraction was found to be easier on a practical level and quicker to conduct. The benefits were immediately recognised as successful primary IMAC was readily achieved. However, the need to restrict expression levels to prevent the toxic effect of periplasmic build up significantly reduced yields. This was compensated for by using larger volume cell cultures, up to 10 litres, to produce sufficient quantities of protein for crystallisation trials.

7.1.2 Structure and function of the Bam system in *C. jejuni*

Although attempts to crystallise and solve the structure of *Cj*BamA experimentally were inconclusive, 3D models were generated using several structural prediction tools. Computational methods used for assessing the dynamic behaviour of the BamA models indicated that the N-terminal region, specifically POTRA domains P1 and P2 of the structure, undergoes a greater range of motion. These findings are attributed to flexibility in the inter-POTRA linker regions and the relative positions of charged patches on the surface of the protein. It is thought that potential energy stored in the structure is transferred to kinetic energy, which is able to drive the assembly mechanism by reducing the barrier presented by the inner leaflet of the outer-membrane. In this model, which is not dependent on the presence of auxiliary components BamBCE, BamD modulates the activity of BamA by binding to, and stabilising the complex structure during assembly (Figure 7.1).

The insertion of β-strands is proposed to begin with the initial binding of the C-terminal end of the cargo to the N-terminal POTRA domain of BamA. A threading analogy is used to describe the proposed motion by which β-strands are inserted, from N- to C-, by inserting two strands at a time. In the proposed sequence, one cycle of the assembly machinery inserts two strands in the first step. The leading strand is pinned at the base of the barrel on the inside of the membrane by a β-hairpin around which the next strand pivots. In the next step the OMP moves slightly into the membrane laterally to free the pore. The next pair is then around the β-hairpin while the POTRA arm is in an extended state. The POTRA domain then compresses to force the next leading strand pair into the BamA pore, which is repeated until the nascent OMP is fully formed.

In this model assembly process, the trailing strand is moved into position in the pore in phase with the lateral-opening while the POTRA arm extends to receive cargo OMP. However, the accessible gate previously seen in published molecular dynamic simulations is slightly shorter than the height of the typical outer membrane protein β-strand (6-8 amino acids in length), suggesting that extension may occur after insertion or perhaps indicating that the BamA opening can open further than seen in their results [39]. The cargo may fold through this opening or the accessibility may be enhanced by a further partitioning of the barrel opening although this has yet to be observed in either experimental studies or simulations. In the model proposed here, POTRAs in the compression phase drag the cargo's next leading strand up to the membrane pore before returning to the extension phase to allow the interior barrel pore again and allow space for the next following strand to enter.

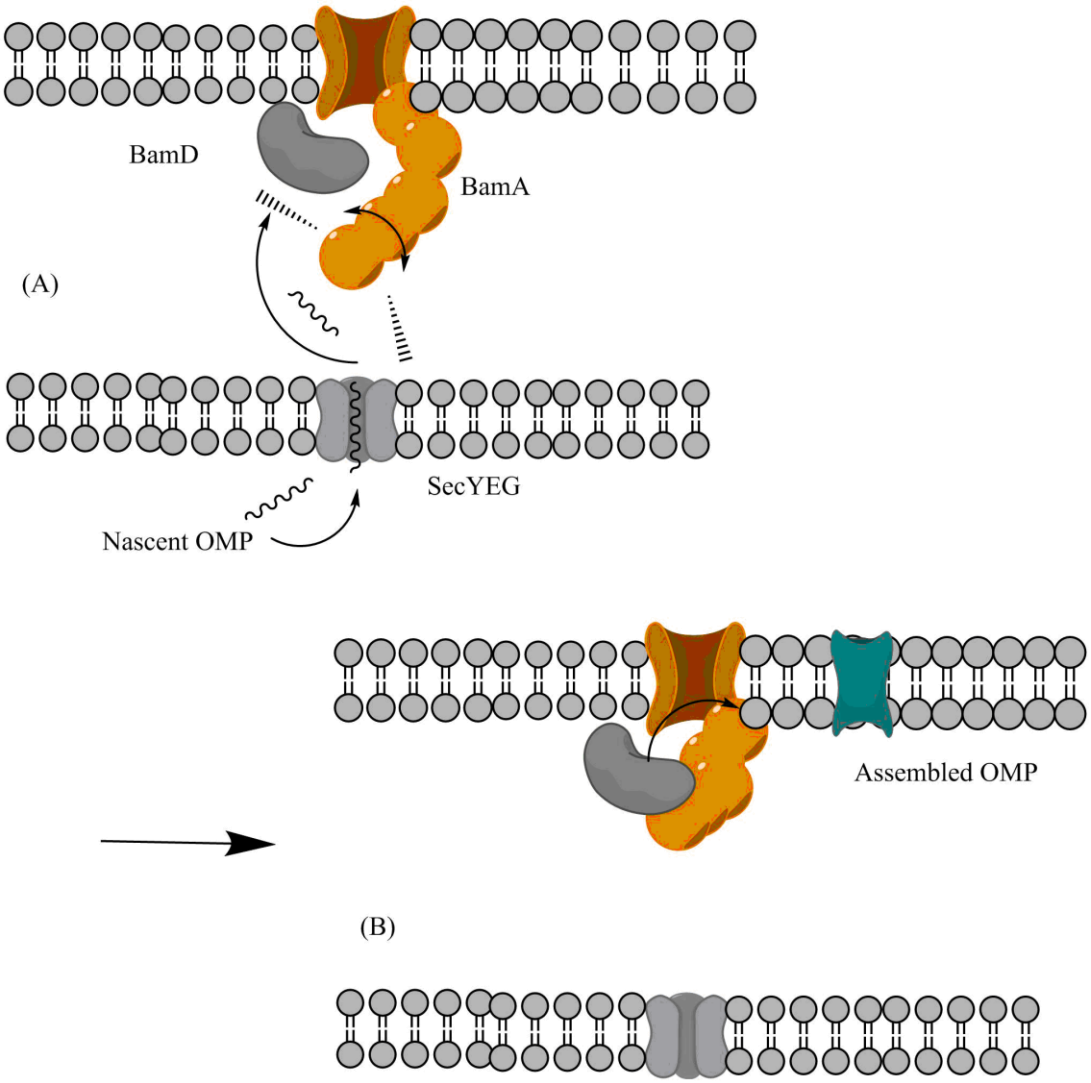


Figure 7.1 Putative mechanism of OMP assembly by BamA and BamD. BamD mediates the dynamic action of the BamA POTRA domain which targets and inserts nascent OMP cargo. In this model, threading of the individual β-strands occurs two at a time a cyclic mechanism which is driven by the hydrophobic effect and regulated through cooperative binding of BamD to the POTRA domains.

7.1.1 Future work

The precise mechanism of β-barrel insertion remains uncertain. To investigate the behaviour and essential dynamics of this protein, molecular dynamics simulations were employed to determine time-dependant motions related to functional attributes important for β-barrel protein assembly in *C. jejuni* and other Gram-negative bacteria. However, the model used was not solved experimentally and further progress towards high-end structural studies, by X-ray crystallography and nuclear magnetic resonance (NMR) is recommended using the sample preparation steps developed here. Furthermore, a wider study of molecular dynamics using experimentally solved structures could be used to investigate the mechanism of β-barrel assembly as well as validate the use of template-based homology modelling programmes in future studies.

It was suggested previously that protein dynamics may occur in a chaotic manner, in which slight changes in the initial environmental conditions result in drastic, unpredictable changes in dynamic behaviour [157]. We have noted here too that none of the simulations ended in the same position, despite the starting conditions being fixed. This was thought to be caused by slight deviations in randomly generated seed velocities, suggesting an inherent level of disorder. Nonetheless, it was observed in these simulations that correlated motion is prevalent in the trajectories of BamA. After the initial results were inspected, it was seen that further high frequency vibrations, analogous to the overtones produced in a sound wave, may be occurring within the structure.

Harmonic mode analysis was therefore conducted to determine if any regular higher frequency motions existed. After detrending the simulations trajectory data to remove the largest component of motion, the patterns clearly indicated correlated harmonic motion occurring within the structural trajectory, which was further shared in all simulation events for both BamA and BamD proteins.

This was done using a detrending and Fast-Fourier transform (FFT) function in R and performed for each 10 ns simulation, along with the extended 100 ns run. Component frequencies, denoted by the harmonics (Hz) (Figure 6.5.1), appear predominantly at 1, 3, 5 intervals, the actual speed of which would be equivalent to 10, 30 and 50 MHz respectively or 2.5 x 10^{-3} meters per second for the slowest motion.

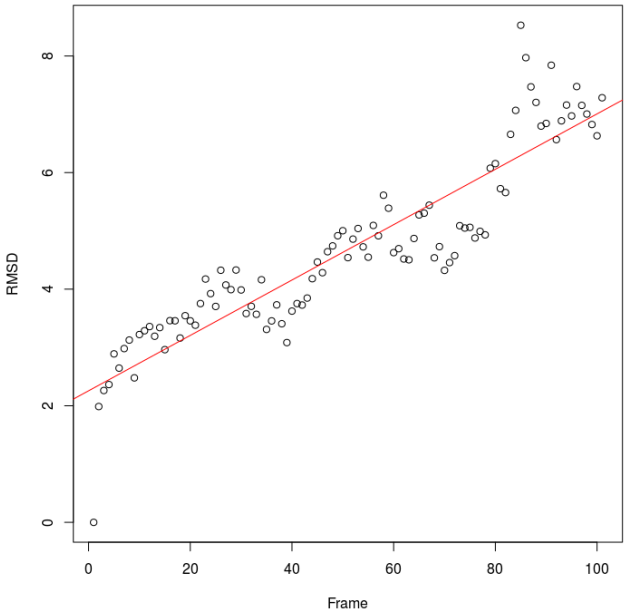


Figure 7.2 Detrended trajectory of 10 ns BamA simulation. Although no experiment data is available to support these measurements, and they lack structural or functional insight, they have the potential to be used in quantitative assessment of results from high-throughput simulation studies and may further serve as output signals for *in silico* protein engineering. Moreover, they indicate that each simulation trajectory holds extractable properties associated with the physical state and dynamics of the protein structure.

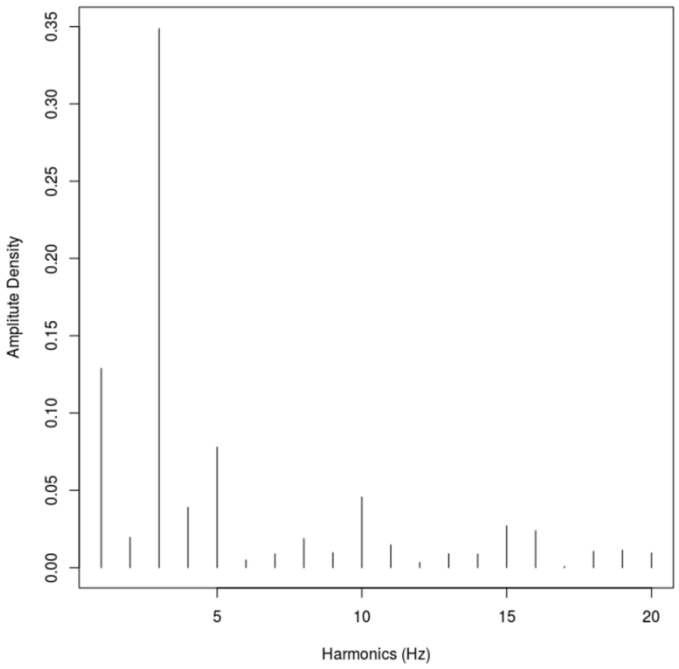


Figure 7.3 Resulting harmonic frequencies of vibrations in a BamA simulation.

Appreciably, there is still a lot that remains to be seen about how the Bam system operates. Specific suggestions for future work that could be taken forward from this thesis:

- Sub-cloning and expression of *Cj*BamD in a pMH1-like *E. coli* strain
- Combined simulations of BamA, BamD and chaperones
- Extended simulations on bigger systems
- Further trajectory analysis to visualise protein behaviour

There is much that can still be attained from molecular dynamics simulations. Because the trajectories are not predictable though, the behaviour of the proteins as observed in these simulations could be described as chaotic. The chaotic nature of protein folding events observed in some simulations could be due to non-linear interactions occurring in the system's force field, physical restraints, and stochastic forces imposed by the system's solvent molecules. Further exploration of these structures could show to what extent the unpredictable behaviour observed in the simulations is caused by the actual behaviour of the proteins themselves [157].

This work could also be used to enhance our understanding of drug-target interactions and to further develop novel therapeutics, especially for the development of new antibiotics. If it is shown that protein modelling and molecular dynamics tools can be used to determine the specific structure and behaviour of homologous protein, then it is feasible that *in silico* drug design programmes could be developed for target unique organisms. This would be highly advantageous as currently the use of broad spectrum antibiotics only accelerates the development of antibiotics resistance in nature. We have shown here though that both *in vitro* and *in silico* research methods can and should be used to better further our understanding of transport processes in Gram-negative bacteria and other biological systems

Acknowledgements

Firstly, I would like to acknowledge Mr. David Griffin who laid the foundation for both the computational and wet lab side of this work and for showing us how to navigate the corridors in the QMC.

I would also like to acknowledge and thank my supervisors at the University of Nottingham, Dr. Neil Oldfield, Dr. David Turner, Prof. Phil Williams and Dr. Boyan Bonev for their continual support, motivation and kindness.

Finally, I would also like to say a big thank you to Dr. Rohanna Hussain and Dr. Isabella Morales for looking after me in Oxfordshire and to my bench supervisor Dr. Chris Grant, Dr. Michael 'Sid' Sadowski, Tim Fell, Sean Ward, Donald, Sarah, Jesus, and the rest of the Antha team for making me feel welcome while on placement in London.

This work was partially funded by a BBSRC Doctoral Training Partnership scholarship, the Diamond Light Source and the Membrane Protein Lab, Imperial College London.

Bibliography

[1] Z. Liu, X. Li, P. Zhao, J. Gui, W. Zheng, and Y. Zhang, "Tracing the evolution of the mitochondrial protein import machinery.," *Comput. Biol. Chem.*, vol. 35, no. 6, pp. 336–40, Dec. 2011.

[2] D. J. Kelly, "The physiology and metabolism of *Campylobacter jejuni* and *Helicobacter pylori*" *J. Appl. Microbiol.*, vol. 90, p. 16S–24S, 2001.

[3] K. K. Nyati and R. Nyati, "Role of *Campylobacter jejuni* infection in the pathogenesis of Guillain-Barre syndrome: An update," *Biomed Res. Int.*, vol. 2013, 2013.

[4] A. Ternhag, A. Törner, A. Svensson, J. Giesecke, and K. Ekdahl, "Mortality following *Campylobacter* infection: a registry-based linkage study.," *BMC Infect. Dis.*, vol. 5, p. 70, 2005.

[5] S. GS and B. MJ, "Fatalities associated with *Campylobacter jejuni* infections," *JAMA*, vol. 253, no. 19, pp. 2873–2875, May 1985.

[6] J. Pritchard, "Guillain-Barré syndrome," vol. 10, no. 4, pp. 399–401, 2010.

[7] R. Koebnik, K. P. Locher, and P. Van Gelder, "Structure and function of bacterial outer membrane proteins: barrels in a nutshell.," *Mol. Microbiol.*, vol. 37, no. 2, pp. 239–253, Jul. 2000.

[8] J. K. Y *et al.*, "A role for the bacterial outer membrane in the pathogenesis of *Helicobacter pylori* infection," vol. 182, no. March, pp. 259–264, 2018.

[9] A. C. Maue *et al.*, "The polysaccharide capsule of *Campylobacter jejuni* modulates the host immune response.," *Infect. Immun.*, vol. 81, no. 3, pp. 665–672, Mar. 2013.

[10] K. T. Young, L. M. Davis, and V. J. DiRita, "*Campylobacter jejuni*: molecular biology and pathogenesis," *Nat. Rev. Microbiol.*, vol. 5, no. 9, pp. 665–679, 2007.

[11] S. Salloway, L. A. Mermel, M. Seamans, G. O. Aspinall, and J. E. N. A. M. Shin, "Miller-Fisher Syndrome Associated with *Campylobacter jejuni* Bearing Lipopolysaccharide Molecules That Mimic Human Ganglioside GD 3," vol. 64, no. 8, pp. 2945–2949, 1996.

[12] T. H. E. C. Veterinary and L. A. Revue, "Bovine Campylobacteriosis : A Review," vol. 22, no. November, pp. 327–330, 1981.

[13] P. H. N. Panzenhagen *et al.*, "Prevalence and fluoroquinolones resistance of *Campylobacter* and *Salmonella* isolates from poultry carcasses in Rio de Janeiro, Brazil," *Food Control*, vol. 61, pp. 243–247, Mar. 2016.

[14] F. Han, S. I. Lestari, S. Pu, and B. Ge, "Prevalence and antimicrobial resistance among *Campylobacter* spp. in Louisiana retail chickens after the enrofloxacin ban.," *Foodborne Pathog.*

Dis., vol. 6, no. 2, pp. 163–171, Mar. 2009.

[15] A. A. Inyinbor, O. S. Bello, A. E. Fadiji, and H. E. Inyinbor, "Threats from antibiotics: A serious environmental concern," *J. Environ. Chem. Eng.*, vol. 6, no. 1, pp. 784–793, Feb. 2018.

[16] C. A. Whitehouse, S. Zhao, and H. Tate, "Antimicrobial Resistance in *Campylobacter* Species: Mechanisms and Genomic Epidemiology," *Adv. Appl. Microbiol.*, Feb. 2018.

[17] L. J. V Piddock, "Quinolone resistance and Campylobacter spp.," *J. Antimicrob. Chemother.*, vol. 36, no. 6, pp. 891–898, 1995.

[18] J. Robin Warren and B. Marshall, "Unidentified Curved Bacilli on gastric epithelium in active chronic gastritis" *Lancet*, vol. 321, no. 8336, pp. 1273–1275, Jun. 1983.

[19] M. M. C. Sheppard SK, "The Evolution of *Campylobacter jejuni* and *Campylobacter coli*" *Cold Spring Harb. Perspect. Biol.*, vol. 22, no. 7, pp. 1–2, Jun. 2015.

[20] J. Parkhill *et al.*, "The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences.," *Nature*, vol. 403, no. 6770, pp. 665–668, Feb. 2000.

[21] L. W. Parfrey *et al.*, "Evaluating support for the current classification of eukaryotic diversity," *PLoS Genet.*, vol. 2, no. 12, pp. 2062–2073, 2006.

[22] P. Kämpfer and S. P. Glaeser, "Prokaryotic taxonomy in the sequencing era - the polyphasic approach revisited," *Environ. Microbiol.*, vol. 14, no. 2, pp. 291–317, 2012.

[23] C. Sohlenkamp and O. Geiger, "Bacterial membrane lipids: Diversity in structures and pathways," *FEMS Microbiol. Rev.*, vol. 40, no. 1, pp. 133–159, 2015.

[24] H. C. Gram, "Über die isolierte Färbung der Schizomyceten in Schnitt- und Trockenpräparaten.," *Fortschritte der Medizin.*, vol. 2, pp. 185–189, 1884.

[25] Y. Li *et al.*, "Correction for Li et al., LPS remodeling is an evolved survival strategy for bacteria," *Proc. Natl. Acad. Sci.*, vol. 109, no. 34, pp. 13877–13877, 2012.

[26] J. Errington, "L-form bacteria, cell walls and the origins of life," *Open Biol.*, vol. 3, no. 1, pp. 120143–120143, 2013.

[27] T. Sikosek and H. S. Chan, "Biophysics of protein evolution and evolutionary protein biophysics," *J. R. Soc. Interface*, vol. 11, no. 100, pp. 20140419–20140419, 2014.

[28] R. S. Gupta, "Origin of diderm (Gram-negative) bacteria: Antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes," *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.*, vol. 100, no. 2, pp. 171–182, 2011.

[29] L. C. S. Antunes *et al.*, "Phylogenomic analysis supports the ancestral presence of LPS-outer membranes in the firmicutes," *Elife*, vol. 5, no. AUGUST, pp. 0–21, 2016.

[30] M. Nakai, "The TIC complex uncovered: The alternative view on the molecular mechanism of protein translocation across the inner envelope membrane of chloroplasts.," *Biochim. Biophys. Acta*, vol. 1847, no. 9, pp. 957–67, Sep. 2015.

[31] R. Aono, N. Tsukagoshi, and M. Yamamoto, "Involvement of outer membrane protein TolC, a possible member of the mar-sox regulon, in maintenance and improvement of organic solvent tolerance of Escherichia coli K-12," *J. Bacteriol.*, vol. 180, no. 4, pp. 938–944, 1998.

[32] S. M. Logan and T. J. Trust, "Outer membrane characteristics of *Campylobacter jejuni*." *Infect. Immun.*, vol. 38, no. 3, pp. 898–906, Dec. 1982.

[33] W. Breyton, C; Haase, W; Rapoport, T; Kuhlbrandt and I. Collinson, "The structure of the bacterial protein translocation complex SecYEG.," *Biochem. Soc. Trans.*, vol. 33, no. Pt 6, pp. 1225–1230, 2005.

[34] J. G. Sklar, T. Wu, D. Kahne, and T. J. Silhavy, "Defining the roles of the periplasmic chaperones SurA, Skp, and DegP in Escherichia coli," *Genes Dev.*, vol. 21, no. 19, pp. 2473–2484, 2007.

[35] D. Gessmann *et al.*, "Outer membrane β-barrel protein folding is physically controlled by periplasmic lipid head groups and BamA.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 16, pp. 5878–5883, Apr. 2014.

[36] M. E. Bauer and W. M. Shafer, "On the in vivo significance of bacterial resistance to antimicrobial peptides," *Biochim. Biophys. Acta - Biomembr.*, vol. 1848, no. 11, pp. 3101–3111, Nov. 2015.

[37] P. Sperandeo *et al.*, "New Insights into the Lpt Machinery for Lipopolysaccharide Transport to the Cell Surface: LptA-LptC Interaction and LptA Stability as Sensors of a Properly Assembled Transenvelope Complex," *J. Bacteriol.*, vol. 193, no. 5, pp. 1042–1053, 2011.

[38] I. E. Gentle, L. Burri, and T. Lithgow, "Molecular architecture and function of the Omp85 family of proteins," *Mol. Microbiol.*, vol. 58, no. 5, pp. 1216–1225, 2005.

[39] J. Bakelar, S. K. Buchanan, and N. Noinaj, "The structure of the β-barrel assembly machinery complex," *Science*, vol. 351, no. 6269, pp. 180–186, 2016.

[40] N. Noinaj *et al.*, "Structural insight into the biogenesis of β-barrel membrane proteins.," *Nature*, vol. 501, no. 7467, pp. 385–390, Sep. 2013.

[41] B. van den Berg, "Going forward laterally: transmembrane passage of hydrophobic molecules through protein channel walls," *Chembiochem*, vol. 11, no. 10, pp. 1339–1343, 2010.

[42] C. Grant *et al.*, "Identification and use of an alkane transporter plug-in for applications in

biocatalysis and whole-cell biosensing of alkanes," *Sci. Rep.*, vol. 4, pp. 1–9, 2014.

[43] P. Z. Gatzeva-Topalova, T. A. Walton, and M. C. Sousa, "Crystal structure of YaeT: conformational flexibility and substrate recognition.," *Structure*, vol. 16, no. 12, pp. 1873–81, Dec. 2008.

[44] P. Z. Gatzeva-Topalova, L. R. Warner, A. Pardi, and M. C. Sousa, "Structure and flexibility of the complete periplasmic domain of BamA: the protein insertion machine of the outer membrane.," *Structure*, vol. 18, no. 11, pp. 1492–501, Nov. 2010.

[45] T. J. Knowles, S. Bobat, M. Jeeves, I. R. Henderson, and M. Overduin, "Secondary structure and 1H, 13C and 15N resonance assignments of the *Escherichia coli* YaeT POTRA domain.," *Biomol. NMR Assign.*, vol. 1, no. 1, pp. 113–115, Jul. 2007.

[46] R. Misra, "First glimpse of the crystal structure of YaeTs POTRA domains.," *ACS Chem. Biol.*, vol. 2, no. 10, pp. 649–651, Oct. 2007.

[47] C. Dong, X. Yang, H.-F. Hou, Y.-Q. Shen, and Y.-H. Dong, "Structure of *Escherichia coli* BamB and its interaction with POTRA domains of BamA.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 68, no. Pt 9, pp. 1134–1139, Sep. 2012.

[48] S. Kim, J. C. Malinverni, P. Sliz, T. J. Silhavy, S. C. Harrison, and D. Kahne, "Structure and function of an essential component of the outer membrane protein assembly machine.," *Science*, vol. 317, no. 5840, pp. 961–964, Aug. 2007.

[49] T. J. Knowles *et al.*, "Fold and function of polypeptide transport-associated domains responsible for delivering unfolded proteins to membranes.," *Mol. Microbiol.*, vol. 68, no. 5, pp. 1216–1227, Jun. 2008.

[50] T. J. Knowles, A. Scott-Tucker, M. Overduin, and I. R. Henderson, "Membrane protein architects: the role of the BAM complex in outer membrane protein assembly.," *Nat. Rev. Microbiol.*, vol. 7, no. 3, pp. 206–214, Mar. 2009.

[51] G. Yinghong *et al.*, "Structural basis of outer membrane protein insertion by the BAM complex," *Nature*, vol. 531, no. 7592, pp. 64–69, Mar. 2016.

[52] D. S. Manning, D. K. Reschke, and R. C. Judd, "Omp85 proteins of Neisseria gonorrhoeae and *Neisseria meningitidis* are similar to *Haemophilus influenzae* D-15-Ag and *Pasteurella multocida* Oma87.," *Microb. Pathog.*, vol. 25, no. 1, pp. 11–21, Jul. 1998.

[53] C. G. Ruffolo and B. Adler, "Cloning, sequencing, expression, and protective capacity of the oma87 gene encoding the Pasteurella multocida 87-kilodalton outer membrane antigen.," *Infect. Immun.*, vol. 64, no. 8, pp. 3161–3167, Aug. 1996.

[54] M. H. Saier, C. H. Ma, L. Rodgers, D. G. Tamang, and M. R. Yen, "Protein secretion and membrane insertion systems in bacteria and eukaryotic organelles.," *Adv. Appl. Microbiol.*,

vol. 65, pp. 141–97, Jan. 2008.

[55] S. Guedin, E. Willery, C. Locht, and F. Jacob-Dubuisson, "Evidence that a globular conformation is not compatible with FhaC-mediated secretion of the Bordetella pertussis filamentous haemagglutinin.," *Mol. Microbiol.*, vol. 29, no. 3, pp. 763–774, Aug. 1998.

[56] S. A. Paschen, W. Neupert, and D. Rapaport, "Biogenesis of beta-barrel membrane proteins of mitochondria.," *Trends Biochem. Sci.*, vol. 30, no. 10, pp. 575–82, Oct. 2005.

[57] L.-X. Shi and S. M. Theg, "The chloroplast protein import system: from algae to trees.," *Biochim. Biophys. Acta*, vol. 1833, no. 2, pp. 314–31, Feb. 2013.

[58] C. T. Webb, E. Heinz, and T. Lithgow, "Evolution of the β-barrel assembly machinery.," *Trends Microbiol.*, vol. 20, no. 12, pp. 612–20, Dec. 2012.

[59] S. Kutik *et al.*, "Dissecting membrane insertion of mitochondrial beta-barrel proteins.," *Cell*, vol. 132, no. 6, pp. 1011–24, Mar. 2008.

[60] J. Bédard and P. Jarvis, "Green light for chloroplast outer-membrane proteins.," *Nat. Cell Biol.*, vol. 10, no. 2, pp. 120–2, 2008.

[61] T. Cavalier-Smith, "Origin of mitochondria by intracellular enslavement of a photosynthetic purple bacterium," *Proc. R. Soc. B Biol. Sci.*, vol. 273, no. 1596, pp. 1943–1952, 2006.

[62] V. Robert, E. B. Volokhina, F. Senf, M. P. Bos, P. Van Gelder, and J. Tommassen, "Assembly factor Omp85 recognizes its outer membrane protein substrates by a species-specific C-terminal motif.," *PLoS Biol.*, vol. 4, no. 11, p. e377, Nov. 2006.

[63] T. Wu, J. Malinverni, N. Ruiz, S. Kim, T. J. Silhavy, and D. Kahne, "Identification of a multicomponent complex required for outer membrane biogenesis in *Escherichia coli*." *Cell*, vol. 121, no. 2, pp. 235–45, Apr. 2005.

[64] J. C. Malinverni *et al.*, "YfiO stabilizes the YaeT complex and is essential for outer membrane protein assembly in *Escherichia coli*." *Mol. Microbiol.*, vol. 61, no. 1, pp. 151–164, Jul. 2006.

[65] R. Voulhoux, "Role of a Highly Conserved Bacterial Protein in Outer," *Science*, vol. 299, no. 262, 2003.

[66] C. L. Hagan, D. B. Westwood, and D. Kahne, "Bam Lipoproteins Assemble BamA in vitro." *Biochemistry*, vol. 52, no. 35, pp. 6108–6113, Sep. 2013.

[67] N. K. Burgess, T. P. Dao, A. M. Stanley, and K. G. Fleming, "Beta-barrel proteins that reside in the *Escherichia coli* outer membrane in vivo demonstrate varied folding behavior in vitro.," *J. Biol. Chem.*, vol. 283, no. 39, pp. 26748–58, Sep. 2008.

[68] P. J. Stansfeld *et al.*, "MemProtMD: Automated Insertion of Membrane Protein Structures into Explicit Lipid Membranes," *Structure(London, England:1993)*, vol. 23, no. 7. pp. 1350–

1361, Jul-2015.

[69] E. B. Volokhina *et al.*, "Species-specificity of the BamA component of the bacterial outer membrane protein-assembly machinery.," *PLoS One*, vol. 8, no. 12, p. e85799, 2013.

[70] T. Sinnige, M. Weingarth, M. Renault, L. Baker, J. Tommassen, and M. Baldus, "Solid-state NMR studies of full-length BamA in lipid bilayers suggest limited overall POTRA mobility.," *J. Mol. Biol.*, vol. 426, no. 9, pp. 2009–21, May 2014.

[71] S. D. Cady, K. Schmidt-Rohr, J. Wang, C. S. Soto, W. F. Degrado, and M. Hong, "Structure of the amantadine binding site of influenza M2 proton channels in lipid bilayers.," *Nature*, vol. 463, no. 7281, pp. 689–92, Feb. 2010.

[72] A. Lange *et al.*, "Toxin-induced conformational changes in a potassium channel revealed by solid-state NMR.," *Nature*, vol. 440, no. 7086, pp. 959–62, Apr. 2006.

[73] K. Nakamura and S. Mizushima, "Effects of heating in dodecyl sulfate solution on the conformation and electrophoretic mobility of isolated major outer membrane proteins from *Escherichia coli* K-12.," *J. Biochem.*, vol. 80, no. 6, pp. 1411–22, Dec. 1976.

[74] M. Renault, M. P. Bos, J. Tommassen, and M. Baldus, "Solid-State NMR on a Large Multidomain Integral Membrane Protein :," pp. 4175–4177, 2011.

[75] S. J. Cordwell *et al.*, "Identification of membrane-associated proteins from *Campylobacter jejuni* strains using complementary proteomics technologies.," *Proteomics*, vol. 8, no. 1, pp. 122–139, Jan. 2008.

[76] A. M. Barry, "Characterisation of the Sec Protein Secretion Pathway in *Campylobacter jejuni*" 2009.

[77] C. J. Oomen, P. van Ulsen, P. Van Gelder, M. Feijen, J. Tommassen, and P. Gros, "Structure of the translocator domain of a bacterial autotransporter," *EMBO J.*, vol. 23, no. 6, pp. 1257–1266, 2004.

[78] B. Clantin *et al.*, "Structure of the membrane protein FhaC: a member of the Omp85-TpsB transporter superfamily.," *Science*, vol. 317, no. 5840, pp. 957–961, Aug. 2007.

[79] E. Callaway, "The Revolution Will Not Be Crystallized," *Nature*, vol. 525, pp. 172–174, 2015.

[80] D. B. Kell, N. Swainston, P. Pir, and S. G. Oliver, "Membrane transporter engineering in industrial biotechnology and whole cell biocatalysis," *Trends Biotechnol.*, vol. 33, no. 4, pp. 237–246, 2015.

[81] R. Tellez and R. Misra, "Substitutions in the BamA β-barrel domain overcome the conditional lethal phenotype of a ΔbamB ΔbamE strain of *Escherichia coli*." *J. Bacteriol.*, vol. 194, no. 2, pp. 317–324, Jan. 2012.

[82] H. W. J. C. Kendrew, G. Bodo, H. M DINTZIS, R. G. PARRISH, "A Three Dimensional Model of The Myoglobin molecule o obtained by X-RAY analysis," *Group*, vol. 181, p. 662, 1958.

[83] M. Caffrey, "Crystallizing membrane proteins for structure determination: use of lipidic mesophases.," *Annu. Rev. Biophys.*, vol. 38, pp. 29–51, 2009.

[84] K. van Holde, *Principles of Physical Biochemistry*. New Jersey: Pearson, 2006.

[85] P. Briggs, M. D. Winn, S. Bailey, and A. Ashton, "Ccp4 Newsletter on Protein Crystallography," *Ccp4.Ac.Uk*, no. 4, 2002.

[86] H. Saibil, "Chaperone machines for protein folding, unfolding and disaggregation," *Nat Rev Mol Cell Biol*, vol. 14, no. 10, pp. 630–642, Oct. 2013.

[87] E. D. Clark, "Protein refolding for industrial processes.," *Curr. Opin. Biotechnol.*, vol. 12, no. 2, pp. 202–7, 2001.

[88] J. Edelman, "The low temperature heat capacity of solid proteins," *Biopolymers*, vol. 32, no. 3, pp. 209–218, 1992.

[89] I. W. Hamley, *Introduction to Soft Matter: Synthetic and Biological Self-Assembling Materials*, 1st (Revis. Chichester: John Wiley & Sons, 2007.

[90] A. M. Seddon, P. Curnow, and P. J. Booth, "Membrane proteins, lipids and detergents: Not just a soap opera," *Biochim. Biophys. Acta - Biomembr.*, vol. 1666, no. 1–2, pp. 105–117, 2004.

[91] V. Postis *et al.*, "The use of SMALPs as a novel membrane protein scaffold for structure study by negative stain electron microscopy," *Biochim. Biophys. Acta - Biomembr.*, vol. 1848, no. 2, pp. 496–501, 2015.

[92] B. J. Berne and R. Pecora, *Dynamic light scattering: with applications to chemistry, biology, and physics*. Courier Corporation, 2000.

[93] A. Meyer, K. Dierks, R. Hussein, K. Brillet, H. Brognaro, and C. Betzel, "Systematic analysis of protein-detergent complexes applying dynamic light scattering to optimize solutions for crystallization trials," *Acta Crystallogr. Sect. FStructural Biol. Commun.*, vol. 71, pp. 75–81, 2015.

[94] G. Binnig and C. F. Quate, "Atomic Force Microscope," *Phys. Rev. Lett.*, vol. 56, no. 9, pp. 930–933, 1986.

[95] K. A. Walther, J. Brujić, H. Li, and J. M. Fernández, "Sub-Angstrom Conformational Changes of a Single Molecule Captured by AFM Variance Analysis," *Biophys. J.*, vol. 90, no. 10, pp. 3806–3812, May 2006.

[96] H. Stahlberg *et al.*, "Two-dimensional crystals: a powerful approach to assess structure, function and dynamics of membrane proteins," vol. 504, no. 3, pp. 166–172, 2001.

[97] A. M. Whited and P. S. H. Park, "Atomic force microscopy: A multifaceted tool to study

membrane proteins and their interactions with ligands," *Biochim. Biophys. Acta - Biomembr.*, vol. 1838, no. 1 PARTA, pp. 56–68, 2014.

[98] M. Shibata, T. Uchihashi, T. Ando, and R. Yasuda, "Long-tip high-speed atomic force microscopy for nanometer-scale imaging in live cells," *Sci. Rep.*, vol. 5, p. 8724, 2015.

[99] D. J. Muller and A. Engel, "Atomic force microscopy and spectroscopy of native membrane proteins" *Nat. Protoc.*, vol. 2, no. 9, pp. 2191–2197, Sep. 2007.

[100] D. J. Müller, C.-A. Schoenenberger, F. Schabert, and A. Engel, "Structural Changes in Native Membrane Proteins Monitored at Subnanometer Resolution with the Atomic Force Microscope: A Review" *J. Struct. Biol.*, vol. 119, no. 2, pp. 149–157, 1997.

[101] D. Small, *Handbook of Lipid Research*. New York: Plenum Press, 1996.

[102] D. A. P. Gutmann *et al.*, "A high-throughput method for membrane protein solubility screening: The ultracentrifugation dispersity sedimentation assay," *Protein Sci.*, vol. 16, no. 12, pp. 1422–1428, 2007.

[103] G. L. Rosano and E. A. Ceccarelli, "Recombinant protein expression in *Escherichia coli*: Advances and challenges" *Front. Microbiol.*, vol. 5, no. APR, pp. 1–17, 2014.

[104] H. C. Hsieh, C. Sheu, F. K. Shi, and D. T. Li, "Development of a titanium dioxide nanoparticle pipette-tip for the selective enrichment of phosphorylated peptides," *J. Chromatogr. A*, vol. 1165, no. 1–2, pp. 128–135, 2007.

[105] N. Dongchun and H. Yihua, "The Expression, Purification, and Structure Determination of BamA from *E. coli*" in *The BAM Complex*, vol. 1329, 2015, pp. 169–178.

[106] M. C. Wiener, "A pedestrian guide to membrane protein crystallization," *Methods*, vol. 34, no. 3, pp. 364–372, 2004.

[107] I. Moraes, G. Evans, J. Sanchez-Weatherby, S. Newstead, and P. D. S. Stewart, "Membrane protein structure determination - the next generation." *Biochim. Biophys. Acta*, vol. 1838, no. 1 Pt A, pp. 78–87, Jan. 2014.

[108] M. R. Rose and T. H. Oakley, "The new biology: beyond the Modern Synthesis" *Biol. Direct*, vol. 2, no. 1, p. 30, 2007.

[109] A. W. BENNETT, "The Origin of Species by means of Natural Selection; or the Preservation of Favoured Races in the Struggle for Life" *Nature*, vol. 5, no. 121, pp. 318–319, 1872.

[110] J. Huxley, *Evolution. The Modern Synthesis*. London: George Alien & Unwin Ltd., 1942.

[111] C. R. Woese and N. Goldenfeld, "How the Microbial World Saved Evolution from the Scylla of Molecular Biology and the Charybdis of the Modern Synthesis," *Microbiol. Mol. Biol. Rev.*, vol. 73, no. 1, pp. 14–21, 2009.

[112] A. Ingles-Prieto *et al.*, "Conservation of protein structure over four billion years.," *Structure*, vol. 21, no. 9, pp. 1690–1697, Sep. 2013.

[113] N. V. Grishin, "Fold Change in Evolution of Protein Structures," *J. Struct. Biol.*, vol. 134, no. 2–3, pp. 167–185, 2001.

[114] D. Stock, "Molecular Architecture of the Rotary Motor in ATP Synthase," *Science (80-.).*, vol. 286, no. 5445, pp. 1700–1705, 1999.

[115] A. P. Hendry and M. T. Kinnison, "Perspective: The Pace of Modern Life: Measuring Rates of Contemporary Microevolution," *Evolution (N. Y).*, vol. 53, no. 6, pp. 1637–1653, 1999.

[116] A. L. Davidson, E. Dassa, C. Orelle, and J. Chen, "Structure, Function, and Evolution of Bacterial ATP-Binding Cassette Systems," *Microbiol. Mol. Biol. Rev.*, vol. 72, no. 2, pp. 317–364, 2008.

[117] R. Schwaab *et al.*, "Haemophilia A: mutation type determines risk of inhibitor formation," *Thromb. Haemost.*, vol. 74, no. 6, p. 1402—1406, Dec. 1995.

[118] C. Leuschner and G. Antranikian, "Heat-stable enzymes from extremely thermophilic and hyperthermophilic microorganisms," *World J. Microbiol. Biotechnol.*, vol. 11, no. 1, pp. 95–114, 1995.

[119] Y. Xia and M. Levitt, "Simulating protein evolution in sequence and structure space," *Curr. Opin. Struct. Biol.*, vol. 14, no. 2, pp. 202–207, 2004.

[120] H. Akashi, N. Osada, and T. Ohta, "Weak selection and protein evolution," *Genetics*, vol. 192, no. 1, pp. 15–31, 2012.

[121] J. Rust, "Chapter 14 Numerical dynamic programming in economics," *Handb. Comput. Econ.*, vol. 1, pp. 619–729, 1996.

[122] I. Eidhammer, I. Jonassen, and W. R. Taylor, *Protein bioinformatics : an algorithmic approach to sequence and structure analysis*. Chichester [etc.]: Wiley, 2004.

[123] D. E. Kim, B. Blum, P. Bradley, and D. Baker, "Sampling Bottlenecks in De novo Protein Structure Prediction," *J. Mol. Biol.*, vol. 393, no. 1, pp. 249–260, Oct. 2009.

[124] R. L. Dunbrack and M. Karplus, "Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains," *Nat. Struct. Biol.*, vol. 1, no. 5, pp. 334–340, 1994.

[125] Y. Zhang, "I-TASSER: fully automated protein structure prediction in CASP8.," *Proteins*, vol. 77 Suppl 9, pp. 100–113, 2009.

[126] A. Fiser and A. Sali, "Modeller: generation and refinement of homology-based protein structure models.," *Methods Enzymol.*, vol. 374, pp. 461–491, 2003.

[127] C. S. Reddy, K. Vijayasarathy, E. Srinivas, G. M. Sastry, and G. N. Sastry, "Homology

modeling of membrane proteins: A critical assessment," *Comput. Biol. Chem.*, vol. 30, no. 2, pp. 120–126, 2006.

[128] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins Struct. Funct. Genet.*, vol. 57, no. 4, pp. 702–710, 2004.

[129] S. A. Adcock and J. A. McCammon, "Molecular dynamics: Survey of methods for simulating the activity of proteins," *Chem. Rev.*, vol. 106, no. 5, pp. 1589–1615, 2006.

[130] I. Bahar, T. R. Lezon, A. Bakan, and I. H. Shrivastava, "Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins.," *Chem. Rev.*, vol. 110, no. 3, pp. 1463–1497, Mar. 2010.

[131] W. Li *et al.*, "The EMBL-EBI bioinformatics web and programmatic tools framework," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W580–W584, 2015.

[132] N. J. Greenfield, "Using circular dichroism spectra to estimate protein secondary structure.," *Nat. Protoc.*, vol. 1, no. 6, pp. 2876–2890, 2006.

[133] N. Sreerama, M. C. Manning, M. E. Powers, J. X. Zhang, D. P. Goldenberg, and R. W. Woody, "Tyrosine, phenylalanine, and disulfide contributions to the circular dichroism of proteins: circular dichroism spectra of wild-type and mutant bovine pancreatic trypsin inhibitor.," *Biochemistry*, vol. 38, no. 33, pp. 10814–10822, Aug. 1999.

[134] E. Minikel,"Circular dichroism on recombinant PrP." *Cam MA,* Dec. 2015.

[135] C. N. Pace, F. Vajdos, L. Fee, G. Grimsley, and T. Gray, "How to measure and predict the molar absorption coefficient of a protein," *Protein Sci.*, vol. 4, no. 11, pp. 2411–2423, 1995.

[136] N. Sreerama, S. Y. Venyaminov, and R. W. Woody, "Analysis of protein CD spectra with a reference protein set based on tertiary structure class," *Biophys. J.*, vol. 80, no. 1, p. 1342, 2001.

[137] "'Estimation of Protein Secondary Structure from Circular Dichroism Spectra: Comparison of CONTIN, SELCON, and CDSSTR Methods with an Expanded Reference Set.'"

[138] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[139] J. Nesper *et al.*, "Omp85(Tt) from *Thermus thermophilus* HB27: an ancestral type of the Omp85 protein family.," *J. Bacteriol.*, vol. 190, no. 13, pp. 4568–4575, Jul. 2008.

[140] A. J. Miles and B. A. Wallace, "Circular dichroism spectroscopy of membrane proteins," *Chem. Soc. Rev.*, vol. 45, no. 18, pp. 4859–4872, 2016.

[141] S. Piana, J. L. Klepeis, and D. E. Shaw, "Assessing the accuracy of physical models used in protein-folding simulations: Quantitative evidence from long molecular dynamics

simulations," *Curr. Opin. Struct. Biol.*, vol. 24, no. 1, pp. 98–105, 2014.

[142] B. Brooks and M. Karplus, "Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 80, no. 21, pp. 6571–6575, Nov. 1983.

[143] A. Hospital, J. R. Goñi, M. Orozco, and J. L. Gelpí, "Molecular dynamics simulations: advances and applications.," *Adv. Appl. Bioinform. Chem.*, vol. 8, pp. 37–47, 2015.

[144] M. E. Tuckerman and G. J. Martyna, "Understanding modern molecular dynamics: Techniques and applications," *J. Phys. Chem. B*, vol. 104, no. 2, pp. 159–178, 2000.

[145] L. Verlet, "Computer Experiments on Classical Fluid" *Physical Review*, vol. 159, no. 1, Jan 1967.

[146] S. Mulholland, E. R. Turpin, B. B. Bonev, and J. D. Hirst, "Docking and molecular dynamics simulations of the ternary complex nisin 2 :lipid II," *Sci. Rep.*, vol. 6, no. 1, p. 21185, 2016.

[147] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, "The MARTINI force field: coarse grained model for biomolecular simulations," *J. Phys. Chem. B*, vol. 111, no. 27, pp. 7812–7824, 2007.

[148] A. Amadei, M. A. Ceruso, and A. Di Nola, "On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins molecular dynamics simulations.," *Proteins*, vol. 36, no. 4, pp. 419–424, Sep. 1999.

[149] R. Elber and M. Karplus, "Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin.," *Science*, vol. 235, no. 4786, pp. 318–321, Jan. 1987.

[150] K. A. Dill, "Dominant forces in protein folding," *Biochemistry*, vol. 29, no. 31, pp. 7133–7155, 1990.

[151] K. A. Dill and J. L. MacCallum, "The protein-folding problem, 50 years on," *Science (80-.).*, vol. 338, no. 6110, pp. 1042–1046, 2012.

[152] W. Kabsch and C. Sander, "Identical Peptide With Different Backbones" *Nature*, vol. 317, no. 6034, p. 207, 1985.

[153] T. Sinnige *et al.*, "Conformational Plasticity of the POTRA 5 Domain in the Outer Membrane Protein Assembly Factor BamA," *Structure*, vol. 23, no. 7, pp. 1317–1324, 2015.

[154] P. J. Fleming *et al.*, "BamA POTRA Domain Interacts with a Native Lipid Membrane Surface," *Biophys. J.*, vol. 110, no. 12, pp. 2698–2709, 2016.

[155] B. D. E. Shaw *et al.*, "Communications of the ACM: Volume 51 Issue 7," *Commun. ACM*, pp. 91–97, 2008.

[156] I. Ohmura, G. Morimoto, Y. Ohno, A. Hasegawa, and M. Taiji, "MDGRAPE-4: a special-

purpose computer system for molecular dynamics simulations," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 372, no. 2021, pp. 20130387–20130387, 2014.

[157] H. Zhou, L. Wang, C. Hill, and N. Carolina, "Chaos in Biomolecular Dynamics," *J. Phys. Chem.*, vol. 100, no. 20, pp. 8101–8105, 1996.

Supplementary material

Table S1. BLOSUM62 residue matching matrix

	C	**S**	**T**	**P**	**A**	**G**	**N**	**D**	**E**	**Q**	**H**	**R**	**K**	**M**	**I**	**L**	**V**	**F**	**Y**	**W**
C	**9**	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	**4**	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	**4**	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	**7**	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	**4**	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	**6**	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	**6**	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	**6**	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	**5**	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	**5**	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	**8**	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	**5**	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	**7**	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	**11**